*ets research institute

April 10
2024

FULL REPORT

Charting the Future of Assessments

AUTHORS

Patrick C. Kyllonen,
Distinguished Presidential Appointee

Amit Sevak,
Chief Executive Officer

Contents

01 Toward the future of assessment: current state and prospects

Contents

Contents

Contents

Contents

Authors and Contributors

Authors

Patrick Kyllonen

Amit Sevak

Co-authors

Teresa Ober

Ikkyu Choi

Jesse Sparks

Daniel Fishtein

Reviewers

Daniel McCaffrey

Rick Tannenbaum

Katherine Castellano

Sandip Sinharay

Randy Bennett

Edmund Gordon

Michael Feuer

James Pellegrino

Narmeen Makhani

Karen Barton

Vanessa Manna

Ido Roll

Contributors

Kadriye Ercikan

Ida Lawrence

Michelle Froah

Sarah Rhame

Christine Betaneli

Kateryna Komarova

Matthew Johnson

Design

&Kind

Editors

Kimberlea Fryer

Ayleen Gontz

SECTION 01.

Toward the future of assessment:

current state and prospects

Introduction

Assessment refers to a broad array of approaches for measuring or evaluating a person's (or group of persons') skills, behaviors, dispositions, or other attributes. Assessments range from standardized tests used in admissions, employee selection, licensure examinations, and domestic and international large-scale assessments of cognitive and behavioral skills to formative K-12 classroom curricular assessments. The various types of assessments are used for a wide variety of purposes, but they also have many common elements, such as standards for their reliability, validity, and fairness—even classroom assessments have standards (Klinger et al., 2015).

We believe the future of assessments will involve a shift in emphasis on what skills will be measured, innovations in how we go about measuring them, the use of advanced technologies for test operations, and an expansion in the value and kinds of information that test-takers will receive from taking the assessment.

In this paper, we argue and provide evidence for our belief that the future of assessment contains challenges but is promising. The challenges include risks associated with security and exposure of personal data, test score bias, and inappropriate test uses, all of which may be exacerbated by the growing infiltration of artificial intelligence (AI) into our lives. The promise is increasing opportunities for testing to help individuals achieve their education and career goals and contribute to well-being and overall quality of life. To help achieve this promise we focus on the evidence-based science of measurement in education and workplace learning, a theme throughout this paper.

A note on tests versus assessments

Throughout this paper we use the terms tests and assessments sometimes interchangeably. The 4th edition of the AERA et al., (1999) *Standards for Educational and Psychological Testing* broadened the definition of test to include assessments; the ETS (2014) *Standards for Quality and Fairness* treats the two terms synonymously. The definitions below are the most current definitions of the two terms from the 5th edition of the Standards (AERA et al., 2014). The definitions are highly overlapping and sometimes synonymous, but assessments is the broader term, allowing for methods other than what are referred to as tests.

Test: An evaluative device or procedure in which a systematic sample of a test-taker's behavior in a specified domain is obtained and scored using a standardized process. (AERA et al., 2014, p. 214)

Assessment: Any systematic method of obtaining information, used to draw inferences about characteristics of people, objects, or programs… sometimes used synonymously with test. (AERA et al., 2014, p. 215)

In this first section of the paper, we review evidence for the value of assessment and discuss how the role of assessment may expand as skills become the new currency. We discuss the many purposes of assessment, from high-stakes examinations and selection tests to low-stakes formative assessments. We review the emerging challenges to testing and assessment, related to perceptions about their value, their focus, validity, fairness and equity concerns. We conclude the first section with a discussion of the prospects for the future of assessments, including the capacity of assessment to provide useful information to test-takers, the importance of identifying key skills and advancing methods for assessing hard-to-measure skills, and the importance of providing opportunities with personalized feedback. The remaining sections of the paper address those themes. The intended audience for this report is broad-ranging from the international scientific community in areas engaged in assessment, particularly education and workforce, to policy-makers and funders in those areas. We try to strike a balance between technical detail and accessibility to the broad audience.

Assessment provides value

Assessment has been with us for centuries and will remain with us. Standardized testing goes as far back as the 3rd century BCE (Wainer, 1987) when Chinese applicants had to pass exams in music, archery, arithmetic, and other subjects to serve as assistants to the Chinese emperor (Himelfarb, 2019). Napoleon revolutionized higher education by founding the École Polytechnique leading to polytechnics across different disciplines by adopting testing and examinations to find talent and avoid nepotism (Bradley, 1975). ETS was founded, as first president Henry Chauncey noted, to find "qualified people from little-known high schools" rather than only those from "blueblood schools" (Lewin, 2002). Today, schools continue to use tests and assessments, but they are used in other sectors as well. Companies use assessments in hiring, leadership development, and for certifying technical skills. Governments and professional associations use assessments in recognition of the need for licensing and certification of skills, especially in key parts of the economy. Throughout history, and around the world—China, Russia, France—assessment has been used for different purposes, as illustrated in the U.S. Congress, Office of Technology Assessment (1992), an important theme for this paper.

The reason testing and assessment have persisted is that it provides value in an efficient and evidence-based way *to support decision-making*. Testing and assessment provide useful information about an examinee's skills to a variety of stakeholders—the test taker, parents, teachers, education administrators, employers, researchers, and policy-makers (Brookhart et al., 2020). A world without testing might instead rely on the "old-boy network" for the decisions that now depend on assessment data. Other methods are problematic. In the U.S. and many parts of the world, grades have become increasingly inflated, particularly in non-STEM fields, and thus provide less information about applicants (Ahn et al., 2019). Non-academic credentials (e.g., resumes; Kessler et al., 2019) are gameable, unfair, and favor the privileged (Chetty et al., 2023). The interview allows gender, racial/ethnic, and physical appearance biases to creep in (Chamorro-Premuzic, 2021). Generative AI products, such as ChatGPT, threaten the validity of college essays, resumes, and other written forms of evaluation as indicators of candidate knowledge, skills, abilities, and experiences. With increasing mobility around the world, and dramatic talent shortages, assessments provide an efficient and economical way to validate skills and knowledge—a nurse from one country can present evidence of qualifying competency in another country. Throughout the history of assessment, there has been a persistent tension between its *equity* and *efficiency* attributes, a theme we return to throughout the paper.

The ETS Human Progress Study (September, 2023; cited in this report as ETS [2023a]) is a set of in-depth interviews with nine world thought leaders on the future, along with a survey conducted in partnership with the Harris Poll. The survey was conducted 18-27 September 2023 with 17,143 respondents, age 18+, from 17 high- and middle-income countries (minimum 1,000 per country) regarding their views on a variety of topics related to the future of assessment and other societal issues and social outcomes. We quote from thought leaders and present survey results from the study throughout this paper.

The data were weighted to ensure representation of the overall population. However, the data might not generalize to the entire country populations and results should be viewed as the opinion of a diverse sample rather than the opinions of populations. Countries were referred to as high-income or middle-income as defined by World Bank definitions.

Assessment provides opportunities, especially to those whose accomplishments or potential would otherwise go unrecognized (Schmill, 2022). Testing provides even more value by giving feedback to the test-taker on where they stand and what they should do next to improve (Wisniewski et al., 2020).

Skills are the future currency

Andreas Schleicher, Director for Education and Skills, and Special Advisor on Education Policy to the Secretary-General at the Organisation for Economic Co-operation and Development (OECD) argues for “skills becoming more like a currency” (ETS, 2023a). The Human Progress Study survey asked several questions on the future of assessment. Table 1 shows that a high percentage of respondents agreed that proof of specific skills will become more important than a university degree and that micro-credentials will become a way to showcase those skills. Not shown here is that agreement was particularly strong among respondents in middle-income countries and the younger cohorts.

TABLE 1.

FUTURE OF ASSESSMENT PREDICTIONS RELATED TO CREDENTIALING

Percentage of respondents who agree with the following statements

	AGREE+ STRONGLY AGREE	STRONGLY AGREE
In the future, proof of specific skills will be more important than a university degree	78%	32%
In the future, micro-credentials (short-term, focused certifications) will become a valuable way to showcase skills.	81%	27%

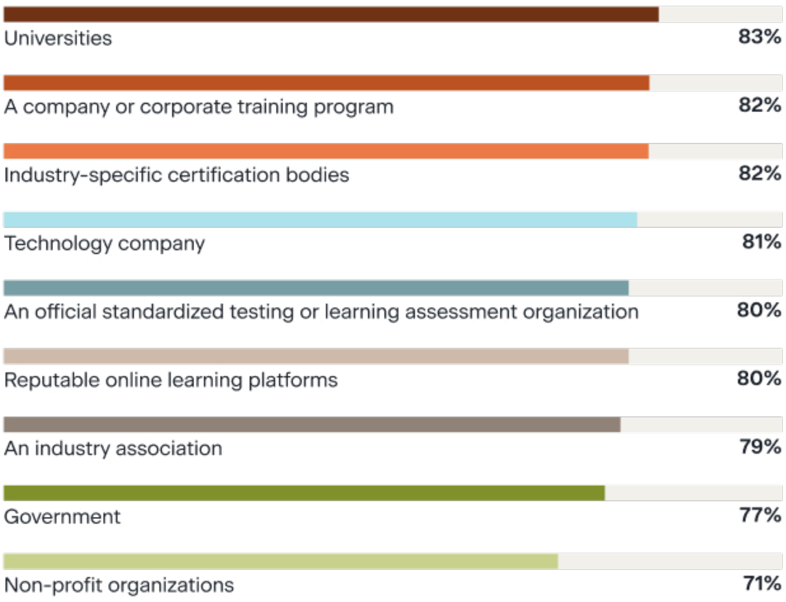
SOURCE. ETS Human Progress Study (September, 2023). Question: "How much do you agree or disagree with the following statements ?" (Strongly disagree/Somewhat disagree/Somewhat agree/Strongly agree)

TABLE 2.

VALUE OF DIFFERENT CERTIFICATION SOURCES

% respondents that agreed with the following statements

SOMEWHAT OR VERY VALUABLE



Survey respondents believe that a variety of certification sources, which will include universities, but also corporate training and testing organizations, will be approximately equally valued in producing certifications and credentials, as shown in Table 2.

This attention to skills and their certification aligns with another theme likely to affect the future of assessments: the growing importance of continuous, lifelong learning. OECD (2021) defines lifelong learning as encompassing “all forms of skill development and knowledge acquisition occurring over the life cycle” (in Section “When does learning occur? The stages of lifelong learning”). As shown in Table 3, respondents widely agreed that continuous learning is the norm, more important now than it ever has been and that it is essential not only for financial stability but for fulfillment and well-being. Employers will continue to invest in ongoing professional development of their employees for productivity enhancement, and that, too, is a part of continuous learning.

TABLE 3.

FUTURE OF ASSESSMENT: IMPORTANCE OF CONTINUOUS LEARNING

Percentage of respondents who agree with the following statements

Future of Assessment: Predictions	AGREE
Continuous learning makes life more fulfilling.	87%
Continuous learning is essential to well-being.	86%
Continuous learning is necessary to create financial stability in today's world.	86%
In a rapidly changing world, continuous learning is now the norm.	86%
Continous learning is more important now than it has been in the past.	85%

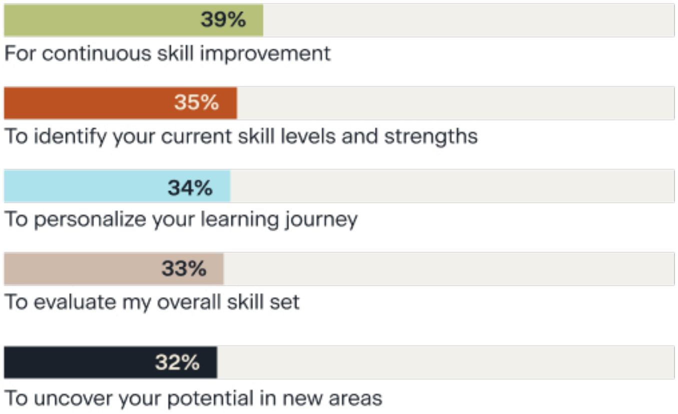
SOURCE. ETS Human Progress Study (September, 2023). Question: "How much do you agree or disagree with the following statements ?" (Strongly disagree/Somewhat disagree/Somewhat agree/Strongly agree)

Testing serves many purposes

Testing is used for many reasons in different contexts. Figure 1 lists the percentage of respondents from the ETS (2023a) Human Progress Study selecting various reasons for taking a test, for reasons other than school admissions and employment screening, which are typically mandated. Reasons range from continuous skill improvement and identifying current skill levels and strengths to uncovering one’s potential in new areas.

FIGURE 1.

PERCENTAGE OF RESPONDENTS INDICATING VARIOUS REASONS FOR TAKING TESTS.



SOURCE: ETS Human Progress Study (2023a); Question: "Which of the following are reasons you would be interested in taking a learning assessment? Select all that apply."

It is important to be mindful of intended use when considering the value of assessment. This principle is enshrined in the *Standards for Educational and Psychological Testing* (AERA et al., 2014), which argued that *validity is the most fundamental consideration in developing and evaluating tests*, with validity defined as “the degree to which evidence and theory support the interpretations of test scores **for proposed uses** of tests” (p. 11; emphasis ours)[1].

An important distinction is between high- and low-stakes uses, which are defined in Table 4, and discussed in the National Research Council (1999a).

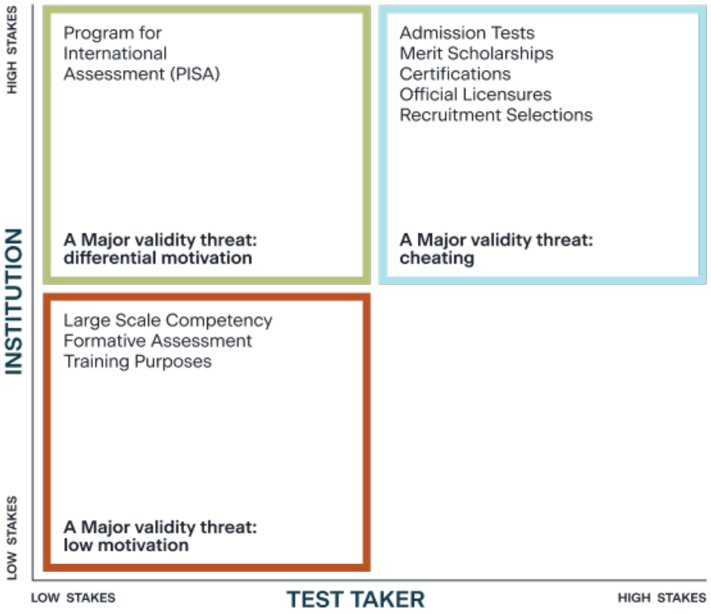


TABLE 4.

DEFINITIONS OF HIGH- AND LOW-STAKE TESTS

High-stakes test:
A test used to provide results that have important, direct consequences for individuals, programs, or institutions involved in the testing.

Low-stakes test:
A test used to provide results that have only minor or indirect consequences for individual, programs, or institutions involved in the testing.

SOURCE: American Educational Research Association et al., 2014, p. 219, 221

While this is a useful binary distinction, Tannenbaum and Kane (2019) following Geisinger (2011) suggest that stakes are tied to the consequences associated with using a test and that there are different kinds and severity of consequences. They argue that for testing applications, such as licensure testing, employment testing, and K-12 accountability testing, one can consider four criteria: positive vs. negative consequences; and the impact, likelihood, and reversibility of the consequences. For example, in a medical licensure test, there are negative consequences for candidates who fail and these consequences are important as they deprive the candidate of practicing, or with a false positive score, they expose the public to an unqualified professional. The likelihood of important consequences is raised for test-takers and the public when scores are near the passing mark and the duration of consequences is the amount of time before a retest is allowed, which could be several months. For employment screening, there may be similar consequences in importance and likelihood, but the duration is less consequential because a candidate can seek another position. However, another aspect of duration is whether the feedback from a test, such as a low score on the honesty dimension in an integrity test, might have longer-lasting consequences on one’s self-esteem, particularly if no guidance is given on how to overcome the perceived deficiency. Tannenbaum and Kane (2019) suggest a “profile of consequences,” that represents a refinement of the high vs. low-stakes dichotomy.

Threats to validity vary depending on whether testing is serving high or low-stakes purposes or the consequences of testing more generally. To take one example, in high-stakes testing, cheating is often a major validity threat. As noted in the definition, the stakes associated with assessments do not necessarily pertain to the one taking the test but can operate on others who have an interest in the results of the testing. Who is most likely to do the cheating is related to who has the most at stake—test-taker, teacher, recruiter, program advocate, policy-maker. In low-stakes testing, motivation or lack thereof is a major validity threat (Wise & Damars, 2005). If a test-taker exerts less than optimal effort due to a lack of incentives or some other

[1] Validity has related but different interpretations in other contexts within and outside testing and psychology generally. From some perspectives validity is a property of the measurement instrument, not the interpretation (see Hood, 1998). See Lissitz (2009) for varying perspectives within educational and psychological measurement.

reason, it is difficult to argue that the test score can be interpreted the same way it would be under optimal effort. Thus, stakes are important when considering the various purposes of testing and assessment. The high-stakes and low-stakes distinction is fundamentally important yet often overlooked[2].

TABLE 5.

EXAMPLES OF USES OF TESTS AND ASSESSMENTS IN VARIOUS SECTORS

Education	Employment	Psychological testing	Program evaluation
Admissions	Prehire:	Psychological conditions diagnosis	Efficacy determination
Formative assessment	Evaluate skills	Cognitive ability assessment	Implementation
Evaluate student learning	Evaluate "intangibles"	Insight into behavior and functioning	determination
Assigning grades	Provide job preview	Determination of values, interests	Formative evaluation
Predict future	Recruit candidates	Preparation of treatment plan	Comparative evaluations
performance Diagnosis (strengths, weaknesses)	Promotion		Program improvement
College credit	Performance appraisal		
Merit award recognition	Provide legal defensibility		
School/District/ Nation monitoring			
Scholarship, internship award			

High-stakes

High-stakes purposes for tests involve using test scores in granting admissions to educational institutions around the world (e.g., the Secondary School Admissions Test [SSAT] for private middle and high school admissions in the U.S.; ETS's GRE® for U.S. graduate admissions; Brazil's Exame Nacional do Ensino Médio [ENEM], national high school exam for degree certification and higher education admissions; China's National College Entrance Examination [Gaokao], given to over 10 million test-takers each year; Japan's National Center Test for University admission; India's Joint Entrance Exam [JEE] for undergraduate engineering programs and National Eligibility cum Entrance Test [NEET] for undergraduate medicine programs; Sweden's Scholastic Aptitude Test [SweSAT]; Australia's Skills for Tertiary Admissions Test [STAT®]), awarding merit scholarships such as ACT and SAT® scores for college, in certification and licensure settings (e.g., ETS's Praxis® Teacher licensure test; Japan's Society of Perinatal and Neonatal Medicine [JSPNM] and Software Testing Qualifications Board [JSTQB]; UK's OSCE, nursing and midwifery licensure exam), in employment recruiting and selection (e.g., SHL Direct, DISC Assessments, Birkman Method, The Predictive Index) and in military personnel selection and classification (e.g., U.S.'s Armed Services Vocational Aptitude Battery [ASVAB]; British Army Recruit Battery [BARB]). Earning certifications to enhance one's resume and job applications and receiving certifications or credentials from an assessment company (see Figure 1) are examples of high-stakes purposes. In-class tests used to determine the grade students receive and whether they pass the course can also be considered high-stakes tests. They can also be high-stakes tests for the teacher or the school, which can incentivize teaching to the test. A potentially high-stakes use for tests is found in the college placement tests category. Placement tests, given to test the academic skills of incoming 2- or 4-year college students in English and math are used to determine whether the incoming student is college-ready and can proceed directly to credit-bearing coursework or must first demonstrate proficiency in remedial courses.

[2] High and low stakes can also be understood as a continuum: a high-stakes test that receives little weight in a decision is not the same as a high-stakes test serving as a sole determinant; and more generally, "important" and "direct" consequences (from the definition) can vary continuously from high to minor and direct to indirect, respectively. Tannenbaum and Kane (2019) provide additional considerations.

However, this is only potentially high stakes because in some cases students can choose to begin college-level courses regardless of their score (Bailey et al., 2010). Advanced Placement® tests (AP®) similarly can be regarded as high stakes in that success can result in college course credit.

The same test can simultaneously be high and low stakes for different involved parties. For example, state accountability tests can be high stakes for parties other than the test-taker, such as the school or district leadership, while simultaneously low-stakes for the student test-taker. Tannenbaum and Kane (2019) provided further considerations for the discussion of testing stakes.

High-stakes testing is particularly vulnerable to Goodhart's (1984) law that "*...when a measure becomes a target, it ceases to become a good measure*." When stakes are high, there is a risk that the test will no longer be a good measure due to corruption pressures. Efforts must be taken to mitigate this risk. One strategy is to caution against overconfidence: despite advances in the science of *measurement* (National Research Council, 2001), a field that might appropriately be thought of as producing *estimates with assumptions*, alternative test score interpretations are generally available.

Low-stakes mixed with high-stakes uses

There are many varieties of low-stakes tests. Figure 1 presents several including ones used for *continuous skill improvement, to personalize one's learning journey, to uncover one's potential in new areas, and to discover career paths aligned with one's strengths*.

Large-scale national educational assessments (e.g., U.S.'s National Assessment of Educational Progress [NAEP]; South Africa's Annual National Assessment [ANA]), and international assessments (e.g., OECD's Programme for International Student Assessment [PISA], Program for the International Assessment of Adult Competencies [PIAAC], and the Study of Social and Emotional Skills [SSES] are low-stakes assessments for the student test-taker, and in some cases also for the teacher, school, and the district, who may complete background or contextual questionnaires. However, the same assessments could be high stakes for the state or national policy-makers and thus results from these assessments can have policy implications, such as responses to the finding of learning loss due to Covid (Mervosh, 2022), or from the presentation of league tables that allow nations and states to see where they stand relative to others and whether they are going in the right direction. "PISA shock" in Germany triggered "...heated public debate and a strong policy response" (Davoli & Entorf, 2018). Results can be used to evaluate environmental effects (e.g., social media; Posso, 2016) or secular trends (e.g., the Flynn effect; Bratsberg & Rogeberg, 2018). These findings indicate that assessments that are low-stakes for the test-taker can have significant, potentially unintended consequences on policy (Feuer, 2012).

Formative assessment, tailoring instruction and providing feedback to students based on their skill levels is another low-stakes use of assessments. Adaptive instruction systems (e.g., Carnegie Learning [BusinessWire, 2024]; Khanmigo [DiCerbo, 2024]) use assessment this way. We review low-stakes uses of formative assessment and feedback in Section 5.

Another low-stakes use is providing normative information to institutions about the skills of their students, or in the case of employers, of their workforce. ETS's Major Field Test line of assessments was designed to provide information to a college about the achievement level of students majoring in a particular field. Results from assessments in different majors, typically resulting from data collection in a capstone course for that major, were used by programs to evaluate program effectiveness and student performance to improve curricula and student outcomes (ETS, n.d.). Similarly, OECD and the EU's Education and Skills Online program was designed to provide information about trainees' literacy, numeracy, and problem solving to diagnose learner strengths and weaknesses and evaluate training against international benchmarks (OECD, n.d.).

One additional assessment use case, which may be high- or low-stakes depending on circumstances, can be found in benchmarking machine capabilities, such as artificial intelligence (AI) progress. For example, the PIAAC assessment was used in a rating study in which AI experts evaluated the degree to which machine algorithms might be able to solve problems appearing on the assessment immediately or in the foreseeable future (Elliott, 2017). Tests similarly have served in AI challenge competitions (Friedland et al., 2004). At one level these are low-stakes applications in that the goal is simply to benchmark, understand, and diagnose machine capabilities. On the other hand, the assessment being administered in high-stakes challenges could provide the typical incentives to game the test, and therefore be considered a high-stakes use.

There are emerging challenges to testing

Despite the many diverse uses of tests and the potential value they provide, the topic of testing has been controversial over the past century (Berman, 2019; Cronbach, 1975; National Research Council, 1999a, 1999b; U.S. Congress, Office of Technology Assessment, 1992) and likely will continue to be. Here we review several emerging challenges that will have to be addressed for assessment to reach its potential in providing positive outcomes for all users.

Concerns that tests do not provide sufficient value

Complaints about standardized testing are long-standing (Grose, 2024). Yet tests can open doors and provide useful information back to the test-takers, policy-makers, and other users of test score information. Consider Table 6, which characterizes agreement levels as statements on the benefits of assessment from ETS (2023a). Over 80% of the respondents agreed that assessments help with finding a job and providing advancement opportunities including equal opportunities regardless of background, boosting self-esteem and career satisfaction, and measuring skills in emerging jobs and roles. This positive sentiment for the value of assessment was particularly true for younger respondents (Gen Z and Millennials), 34% to 40% of whom indicated “strongly agree” with these assessment benefits.

TABLE 6.

PERCEIVED BENEFITS OF ASSESSMENTS

Learning assessments can...	AGREE	STRONGLY AGREE
help individuals to achieve better job opportunities and career advancement.	85%	40%
contribute significantly to boosting individual self-esteem.	84%	37%
can contribute significantly to boosting overall career satisfaction.	84%	38%
provide valuable opportunities for advancement.	84%	34%
effectively measure skills relevant to emerging industries and job roles.	83%	35%
bridge the skills gap to provide equal opportunities for advancement (e.g., across different backgrounds such as socioeconomic, racial, gender, etc.).	82%	34%

SOURCE. ETS Human Progress Study (September, 2023). Question: "How much do you agree or disagree with the following statements?" (Strongly disagree/Somewhat disagree/Somewhat agree/Strongly agree) Note: *Column "agree" is overall; "strongly agree" is Gen Z and Millenials only, strongly agree is approximately 10%-20% lower for Gen X and Boomers.

But testing also requires an investment on the part of the test-taker and those who support the test-taking activity. The investment is in preparation and testing time and effort and potential reputation risk. *An at least implicit cost-benefit calculation takes place to justify the time and effort expenditure by all involved parties.* The more value a test provides to the test-taker and supporters, the more justified the effort and investment will be. Thus, it is important that testing provides value back to the test-taker and stakeholders, a Return on Test (ROT), that justifies the expense.

Tests often fail to provide useful, actionable feedback; they fail to provide users insights that might help determine the next steps to achieve education and career goals. The future of assessments will largely be concerned with providing useful information back to the key stakeholders, especially test-takers, to change the cost-benefit ratio of testing for test-takers and all concerned. Tests will shift from what one knows now to what one can do with the information provided by the test—providing a recommendation path forward. We address these issues in Section 5.

Concerns about the focus of tests being too narrow

A common argument for testing is that we measure what matters, so testing signals our values. *But too often the reverse is true, that we elevate the importance of whatever it is we happen to be testing*. Schrum and Levin (2012) argued that we too often restrict the meaning of exemplary schools to those producing high achievement test scores, which misses a much broader set of skills that contribute to educational attainment and economic outcomes. That is, the focus of tests traditionally has been overly narrow, perhaps at least partly due to focusing on what is easy to measure rather than on what is most important. Educational attainment and workforce and life success require the development of skills beyond those that can be easily measured by mathematics and language tests. It is critical for the future of assessments to identify the most important skills for education, the workforce, and life, and to develop valid and reliable methods for assessing those. We address these issues in Section 2.

Concerns about validity and lack of trust in the scores

Tests do not always measure the skills they purport to measure. For example, in low-stakes settings, students can be unmotivated and disengaged, and then scores from the test are not useful indicators of what students know and can do. For example, we compare states' and nations' achievement levels with large-scale assessments but do not account for differences in effort that might be partly responsible for those differences, despite knowing that effort differences do affect test scores (Liu et al., 2012). Another reason why a test may fail to provide an accurate picture of a student's skill levels is due to cheating or having experienced instruction teaching directly to the test. A more general concern here is with a lack of security around the testing process allowing scores to overstate the skill levels of test-takers or school quality. A third complaint concerns weak testing methods, such as self-reports, particularly for skills and constructs that are hard to measure (Stecher & Hamilton, 2014). For example, persistence and curiosity may be important qualities for students, but if the assessment relies exclusively on self-reports, that might cause those who are interested in such student qualities to lose confidence in conclusions about them drawn from the assessment. The future of assessments is likely to be concerned with better measures of hard-to-measure skills. We address these issues in Section 3.

Concerns about fairness and equity

A major concern many have about testing is that tests are not fair and equitable for all test-takers, resulting in a general lack of trust in the scores, and leading to oppositional attitudes towards testing itself. From this perspective, tests can fail to accurately measure skills if the test-taker is different from the test designer, with respect to culture, gender, language, disability status, or socioeconomic status. More generally, Solano-Flores (2019) argued that tests are cultural products and therefore require consideration of a broad variety of culturally related issues as part of the validity argument for a test.

In addition, tests may be viewed as inequitable because they fail to level the playing field and fail to account for past inequities which might reflect differential opportunities to learn (Darling-Hammond, 2001). As a result, according to this view, tests do not support those so disadvantaged and instead might contribute to inequity, growing inequality, and polarization (see Herman et al., 2023 for an introduction to a special issue of *Educational Assessment* concerned with these issues; and Bennett's, Solano-Flores', and Randall's reflections and recommendations from the same issue). These real or perceived barriers in testing can be exacerbated when looked at in a global context or for students or workers from one country being assessed on standards from another country or culture, such as an Asian worker seeking employment in the United States.

Issues regarding fairness in testing are addressed in the *Standards for Educational and Psychological Testing* (AERA et al., 2014) as well as in the *ETS Standards for Quality and Fairness* (ETS, 2014) and *ETS Guidelines for Developing Fair Tests and Communications* (ETS, 2022)—and other similar documents.[3] Although standards from such documents do not automatically translate to practice (Solano-Flores, 2023), nor necessarily have the statutory backing to be used exclusively in fairness legal defenses, they nevertheless are considered "widely applicable advisory sources" (Biddle & Nooren, 2006, p. 219). As noted in AERA et al., (2014, p. 2):

> ...although the *Standards* is not enforceable by the sponsoring organizations, it has been repeatedly recognized by regulatory authorities and courts as setting forth the generally accepted professional standards that developers and users of tests and other selection procedures follow. Compliance and

[3] Testing fairness issues are also considered in International Test Commission (2001, 2013, 2017), International Test Commission and Association of Test Publishers (2022), and Society for Industrial and Organizational Psychology (2018).

> noncompliance with the *Standards* may be used as relevant evidence of legal liability in judicial and regulatory proceedings. The *Standards* therefore merits careful consideration by all participants in the testing process.

The *Standards* (AERA et al., 2014) consider fairness to be "an overriding, foundational concern" and a "fundamental validity issue" requiring "attention throughout all stages of test development and use" (p. 49). It also advocates for the "fair and equitable treatment of all test-takers during the testing process." The *Standards* also argue that "…a prime threat to fair and valid interpretation of test scores comes from aspects of the test or testing process that may produce a construct-irrelevant variance in scores that systematically lowers or raises scores for identifiable groups of test-takers and results in inappropriate score interpretations for intended uses." It suggests that construct-irrelevant components can be introduced by inappropriate sampling of test content, lack of clarity in test instructions, unnecessary item complexities, scoring criteria that may favor one group, and that "…opportunity to learn … can influence the fair and valid interpretations of test scores for their intended uses" (p. 54).

A major challenge for the future of assessment will be to address the fairness and equity issues articulated here. ETS (2014, 2022) has developed fairness guidelines for both tests and communications generally that provide concrete guidance on addressing the fairness issues raised in the *Standards* (AERA et al., 2014). ETS (2022) presents four fundamental principles: (a) measure the important aspects of the intended construct; (b) avoid construct-irrelevant barriers to the success of test-takers; (c) provide assessment design, content, and conditions that help diverse test-takers show what they know and can do so that valid inferences are supported; and (d) provide scores that support valid inferences about diverse groups of test-takers. ETS (2022) follows these with specific guidelines to support the general principles.

In addition to concerns about test fairness, there are concerns about equity. Between-group disparities in test performance might reflect at least in part differential opportunities to learn, and tests can help identify opportunity gaps (National Academies of Science, Engineering, and Medicine, 2019). However, the view that tests propagate injustice itself is being challenged by the reenergized view that tests reflect students' academic achievements regardless of background, with greater predictive accuracy than grades and other measures considered in admissions, thus providing opportunities to low-income and underrepresented minority applicants (Deming, 2024; Flanagan, 2021; Leonhardt, 2024; McWhorter, 2024). Furthermore, tests can serve as a form of instruction and thereby address equity issues; we expect that a major focus in the future of assessments will be to develop ways to accomplish assessment *for* education, "addressing equitable opportunities to learn" (The Gordon Commission, 2013, p. 150). We address the issue of tests providing feedback in Section 5.

Future of assessment prospects

In this paper, we address the challenges and concerns identified in the previous subsection and argue that an overarching theme for the future of assessments is that assessments will be skills-based, technology-enhanced, and led by developments in AI and related technologies. In recognition of its role in promoting learning, future assessment *will be less deficit-focused, guiding learners to build on their strengths to help them achieve their education and career goals*. Negative feedback is particularly detrimental to the motivation and performance levels of low-power individuals, ones who have less ability to control resources (Straub et al., 2023). Future assessments will provide feedback that is actionable and centered on the test-taker and the concrete actions the test-taker can take.

Providing useful information to test-takers and stakeholders

Future assessments should strive to provide useful, easy-to-understand, reliable, valid, fair, and trustworthy (based on a secure process) information to test-takers and other stakeholders. Assessments should be cost-effective, in relevant languages, and where possible draw insights or be actionable. That information can take the form of certifications, scores, badges, and other indicators of where test-takers stand on the skills most critical for further education and for the current and future workforce, along with actionable feedback that provides information to test-takers and stakeholders on how the test-taker can achieve their educational and career goals.

Identification of key skills

To provide useful information to test-takers it is necessary to identify the most important skills needed for attaining one's education and career goals. Identifying key skills will require compiling evidence about the future viability of skills using various methodologies—surveys, job trends, financial scans—to help determine which skills will increase in importance and which skills will become obsolete. Conducting such analyses (e.g., Autor et al., 2024; Eloundou et al., 2023; Frey & Osborne, 2017; Lassébie & Quintini, 2022) may enable the production of metrics that will help us place investment bets—to determine which skills to invest in for schools, the workforce, and society.

Advancing methods for assessing hard-to-measure skills

Many skills that are increasingly important today and are likely to grow in importance in the future are hard-to-measure skills, such as communication, creativity, and collaboration (see Table 5). Because they are hard to measure, yet important to measure, we tend to use simple self-reports and others' evaluations to measure them. But these methods are not as powerful as the methods we use to measure technical skills, or the so-called hard skills, like mathematics and reading. Self- and others' reports will remain in use, but they are associated with well-documented biases, such as response style (the tendency to respond in a similar way regardless of the construct being measured; He et al., 2014), halo (the tendency to rate a target the same way regardless of the attribute on which the target is being measured; Cooper, 1981), and reference bias (the tendency for respondents to use different standards in ratings; Lira et al., 2022). To supplement or replace these measures, there is a need to develop engaging, personalized, and contextualized performance tasks including games, simulations, and interactive and collaborative tasks, which do not depend on subjective ratings. A trend for the future of assessments might be to move away from overly standardized approaches to ones that can better be characterized as "personalized, differentiated, adapted, culturally and linguistically relevant, and context-based" (Morell, 2015, p. 2). Sireci (2020; from section, *Understanding Understandardization*) argues that understanding personal characteristics that might interact with the conditions of testing and accommodating such personal characteristics might "lead to more accurate interpretations of students' true proficiencies."

The future of assessment will also involve the development of methods to measure naturally occurring behaviors. This will include the analysis of process data, the keystrokes, conversations, response times, and other learning and performance indicators that can be used to draw inferences about the course of development and the status of skills.[4] Note that these methods could be applied to any type of skill—affective, behavioral, or cognitive (ABC; Liu, Kell, et al., 2023).

An important part of this effort will be to develop metrics for evaluating whether and the degree to which we have been successful in devising new measures. We can rely on traditional psychometrics metrics, including validity, reliability, fairness and equity. We also can evaluate success by the degree to which our efforts are valued and align with key and expanding markets.

Providing opportunity to test-takers and other stakeholders with personalized feedback

Providing useful feedback to test-takers will require the identification and implementation of learning principles emanating from multiple disciplines along with efficacy evidence. These disciplines include educational, cognitive, and industrial-organizational psychology; the learning sciences; and neuroscience. Application domains such as human factors, training, and human-computer interaction and instructional domains such as computer-supported collaborative learning and adaptive learning or intelligent tutoring systems also can provide findings and principles that can be incorporated into testing practice. Important research in the fields of AI and education (Koedinger et al., 2023; Zapata-Rivera & Hu, 2022) can inform how testing can provide useful information to test-takers to enhance their learning and the benefits they receive from testing. Providing feedback should also be an ongoing process. Eighty-seven percent of the respondents in ETS (2023a) agreed that "…learning assessments should provide ongoing feedback, not just a one-time snapshot of performance." The benefits of feedback should not be limited to the test-taker. Policy-makers, teachers, and other stakeholders can also benefit from informative, actionable feedback.

Providing feedback shares similarities with educational and health interventions of all kinds and therefore can draw lessons from those fields. For example, the broader field of *implementation science* is described as "…the scientific study of methods to promote the systematic uptake of research findings and other evidence-based practices into routine practice, and, hence, to improve the quality and effectiveness of health services" (Bauer et al., 2015) and may provide useful guidance for how the administration of feedback can improve learner outcomes. Lessons can also come from improvement science, which is designed to

[4] Test-less assessments (i.e., assessments measuring naturally occurring behaviors) has become a highly charged topic. The 2024 EU AI Act https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf prohibits the "use of AI systems to infer emotions of a natural person in the areas of workplace and education institutions" (p. 108), with safety and medical exceptions, and cautions that "AI systems providing social scoring of natural persons by public or private actors may lead to discriminatory outcomes and the exclusion of certain groups" (p. 26).

accelerate learning-by-doing, guiding the development and refinement of new tools and processes (Hinnant-Crawford, 2020).

Themes for the future of assessment and organization of the paper

We believe it is useful to organize the future of assessments around themes, which will be covered over the next four sections. Themes reflect distinct bodies of work and scientific background, often covered in non-overlapping literature. Yet, advances on all fronts represented in the themes are essential for the future of assessment. The following section, *Section 2: Skills for the future: effects of technology advances,* is largely based on studies in economics and artificial intelligence. *Section 3: Innovative measures: new approaches for assessing hard-to-measure skills* draws from diverse fields—cognitive psychology, industrial-organizational psychology, personality psychology and others. *Section 4: Operations breakthroughs: AI and technology-enabled advances* largely reflects the traditional concerns of the testing industry in developing, scoring, and reporting on tests, and draws from educational measurement and psychometrics, operations research, artificial intelligence, and others. *Section 5: Feedback: learning science-driven insights and action plans for test-takers* draws from cognitive psychology, educational psychology, the learning sciences, and adaptive instruction. We conclude with Section 6: Summary and conclusions.

SECTION 02.

Skills for the future:

effects of technology advances

We believe that the future of assessments will largely be driven by the skills needed to be a productive, informed citizen, to maintain health and well-being, and to be a contributing member of the community and society. Entrepreneurs will want those skills and employers will seek those skills in hiring new workers and employers will develop those skills for the current workforce. Higher education will respond to the demand for skills in their offerings and in the popularity of majors and other forms of certification and recognition. K-12 education will follow suit and evolve standards and curricula—the growth in standards, curricula, and assessments for social emotional learning (SEL) skills over the past decade is a case in point (Burrus et al., 2022). Government and industry will develop research and development agendas to ensure the development of those skills. In all cases, educators and employers will want to know about students', applicants', and incumbents' skills to be able to make good decisions in admissions, hiring, promotion, and student and workforce development. This has been the role for assessments and will continue to be.

What might be different today is the rapid pace of change brought on by developments in technology and AI. There is no shortage of future projections that forecast an overhaul of the economy and the nature of future work and occupations, which will be calling for skills that are not yet foreseen. A recent study by Dell (2018) found that 56% of the 3,800 global business leaders they surveyed speculated that "schools will *need to teach how to learn* rather than what to learn to prepare students for jobs that don't yet exist." In ETS's (2023a) Human Progress Study, respondents indicated many areas in which AI is likely to have an impact, on the need to update skills, to combine technical and human skills, and to identify skills for new job opportunities that do not yet exist (Table 7). As Eric Lavin, Partner, Avalanche VC noted in the ETS Human Progress Study, "Learning how to learn is probably the key skill. The half-life of skills is getting shorter as more and more technology comes in. The most important ability is learning how to use the new tools in a way that is resonant with being a human and the job to be done."

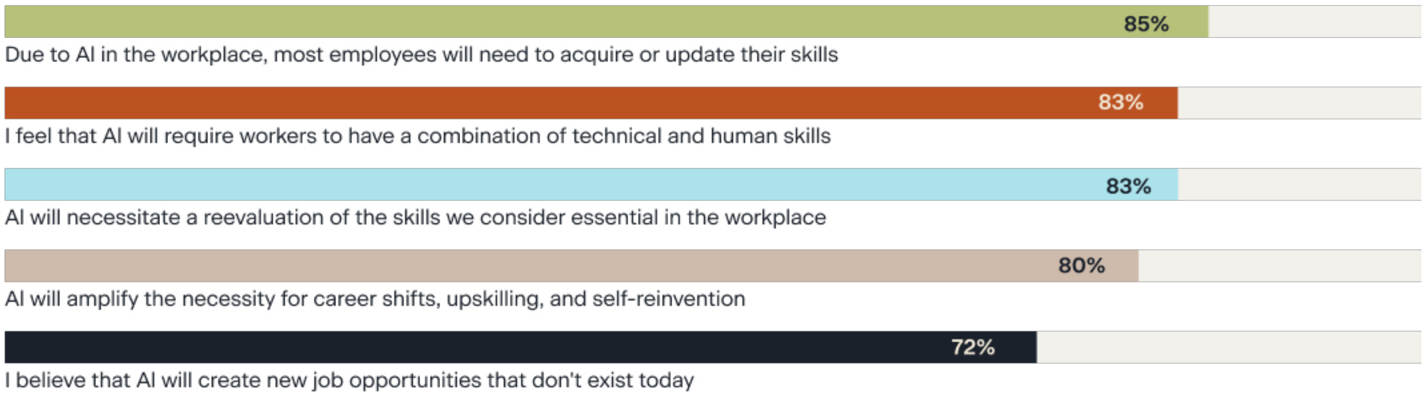
> "The biggest mismatches are now on the quality and relevance of skills."

**Andreas Schleicher,
Director for the Directorate
of Education and Skills OECD**

TABLE 7.

IMPACT OF AI ON THE FUTURE OF SKILLS

% respondents that 'agree' and 'strongly agree' with the following statements



SOURCE. ETS Human Progress Study (September, 2023). Question: "How much do you agree or disagree with the following statements?" (Strongly disagree/Somewhat disagree/Somewhat agree/Strongly agree)

This section is organized as follows. First, we review the skills in demand today, based on respondent perceptions from the ETS (2023a) Human Progress Study, employer and educator surveys, analyses of job ads, and trends in research, policy, and practice. Next, we note that the most highly sought skills today are not the traditional curricular skills that have been the focus of assessment attention for the past 100 years, but are hard-to-measure skills, which presents an assessment challenge. Next, we discuss the skills that are likely to be in demand in the future, based on analysis of trends and based on analyses of the effects of AI and new technologies on the changing nature of skills. We conclude with a discussion of the implications for the future of assessment.

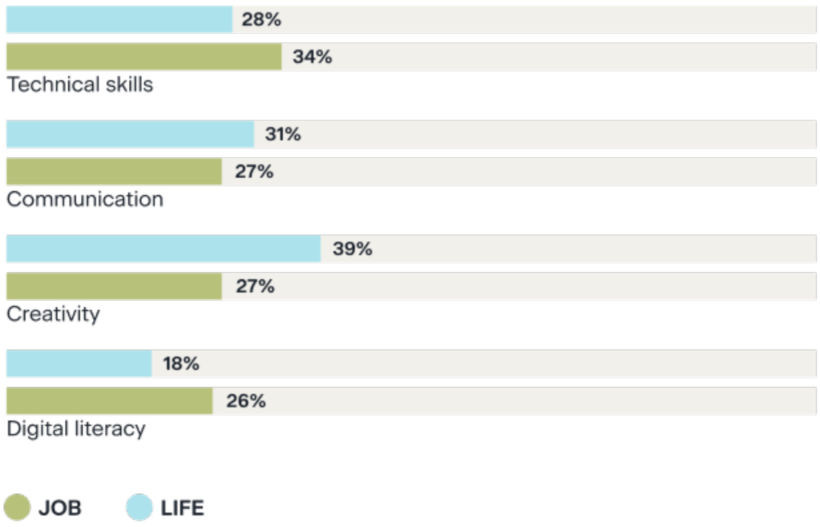
Skills in demand today

We examine evidence for the importance of skills in three sectors—the workforce, higher education, and K-12 education. The skills demanded and developed in the different sectors may be different and the methodology for identifying those skills is different across sectors.[5] Figure 2 indicates what

FIGURE 2.

SKILLS AS NEEDED FOR THE JOB MARKET OR LIFE SUCCESS.

% of respondents selecting various skills as needed for the job market or life success



Source: ETS Human Progress Study (2023a); **Questions:** "What skills do you think you will need to acquire (or improve) to be competitive in the job market in the next 2-3 years? Please select up to three." 11% selected N/A. "What types of skills are most necessary for success in life? Please select up to three." 3% selected N/A

[5] Although it is beyond the scope of this article, there are important additional considerations in determining future skills priorities, such as the geography of skills (Moro et al., 2021).

skills respondents in the ETS Human Progress Study indicated they believed were needed for the job market or for life success. For the job market, technical skills led, followed by creativity, communication, and digital literacy. For life success, communication and problem-solving led, followed by creativity, technical skills, time management, and perseverance.

Skills employers are looking for

Various approaches have been taken to determine the kinds of skills employers are looking for in new hires and seek to develop in their current workforce. Employer surveys reflect what employers say they want in the way of the skills employees ought to have; the

TABLE 8.

TOP-RATED SKILLS BASED ON EMPLOYER SURVEYS AND ANALYSES OF JOB ADS

TOP-RATED SKILLS BASED ON EMPLOYER SURVEYS

NACE 2022

- Problem-solving skills
- Ability to work on a team
- Strong work ethic
- Analytical and quantitative skills
- Communication skills
- Technical skills

WILKIE (2019)

- Communication skills
- Listening skills
- Critical-thinking skills
- Interpersonal skills

WEF 2023

- Analytic thinking
- Creative thinking
- Resilience, flexibility, & agility
- Determination of values, interests
- Curiosity & life-long learning
- Technology literacy
- Dependability & attention to detail
- Empathy & active listening
- Leadership & social influence

TOP-RATED SKILLS FORM ANALYSES OF JOB ADS

Rios et al. (2020)

- Oral & written communication
- Collaboration
- Problem solving
- Communication
- Social intelligence
- Self-direction

Shafer et al. (2023) — Geoscientists

- Written communication

Mankii (2023) — Teacher Educator

- Interpersonal skills, teamwork
- Sensitivty to cultural diversity
- Intercultural understanding
- Professionalism
- Leadership skills

Burning Glass Baseline Skills*

- Communication skills
- Teamwork and collaboration
- Organizational skills
- Problem solving

*Burning Glass divided skills into 3 categories--technical skills, software skills, and 142 baseline skills. Only a summary of the baseline skills is reported here.

analysis of job ads identifies the skills employers are currently hiring for, which should be consistent with survey responses, but does not have to be. It is useful to examine both. Table 8 presents the results from several studies using these approaches. The studies are summarized in the sections following the table.

It can be seen that communication skills, including listening skills, are highly rated across studies, as are the clusters of *cognitive skills*, such as critical thinking, analytical thinking, technical skills, and problem-solving; *interpersonal skills*, such as teamwork, collaboration, social intelligence, and social skills; and *intrapersonal skills* such as work ethic, organizational skills, self-direction, motivation, and self-awareness. There are also mentions of cultural skills (sensitivity to cultural diversity; intercultural understanding). It is also noteworthy that there is considerable overlap in findings across the different studies when allowances are made for variations in terminology.

These lists do not reflect all that is important in the future of skills, particularly because they focus on the perspective from the workforce, and do not reflect important life skills. OECD (2015) argued for the importance of aspects of life such as health, family life, civic engagement (OECD, 2023), and life satisfaction on individual well-being and social progress. Nor do these lists capture the important knowledge taught in formal education as part of the curriculum, such as mathematics, language, and science skills. Nevertheless, the lists above provide an important set of skills for which assessment is not yet fully developed and that therefore represents an opportunity for future growth. The following two sections discuss this summary of skills in more depth.

Employer surveys

Employer surveys poll business leaders on what they look for on candidates' resumes, what skills they might believe are important, and related topics. It is important to acknowledge the limitation of these surveys: samples are not random and are often small and subject to response bias, and question wording can affect answers. Still, employer surveys provide some evidence for employer preferences for different skills in the workforce.

The National Association of Colleges and Employers' (NACE) has been conducting annual surveys of U.S. employers for the past decade or so. In a recent Job Outlook report (National Association of Colleges and Employers, 2022) they found that the top attributes employers look for on resumes were *problem-solving skills, ability to work on a team, strong work ethic, analytical and quantitative skills, communication skills, and technical skills,* with 50 to 61% of employers rating these as *very* or *extremely important*. They also found that the percentage of employers who screen candidates by grade-point average dropped precipitously from 73% to 37% in the past four years, perhaps indicating the decreased employer attention given to curricular achievements and the increased attention to what are sometimes referred to as durable skills. A Cengage/Morning Consult survey of 650 employers (Cengage, 2019) found that the top skills in demand were communication skills, listening skills, critical-thinking skills, and interpersonal skills, with the percentage of employers saying these skills were "very important" to gaining leadership positions at their organizations ranging from 73% to 77%. Results from these two studies essentially duplicate findings from a decade and a half earlier (Casner-Lotto & Barrington, 2006), which asked employers how important to success various skills were and found that *applied skills* were more likely to be rated "very important" than *basic knowledge/skills*, such as math, reading, and science. Applied skills included professionalism/work ethic, teamwork/collaboration, critical thinking/problem-solving, oral and written communications, diversity, and leadership.

To supplement the U.S. surveys, a survey by the World Economic Forum (Di Battista et al., 2023) studied employers outside the U.S., asking them about the core skills today. Employers identified *analytic thinking; creative thinking; resilience, flexibility, and agility; motivation and self-awareness; curiosity and life-long learning; technology literacy; dependability and attention to detail; empathy and active listening;* and *leadership and social influence* as the most highly rated skills, with 39-67% of the respondents rating them as core skills. Employers also were asked about the future (five years from now) and *analytic and creative thinking remained on top, but curiosity and lifelong learning, plus technology literacy* increased substantially to join *resilience, flexibility, and agility* as the predicted top five skill clusters for the future.

Analysis of job ads

Analyzing job advertisements, the means by which employers recruit workforce entrants, should supplement employer surveys in helping to determine the value of various skills in the workforce. Rios et al., (2020) examined 142,000 job advertisements from an employment website for job listings and found 70%

of job ads requested “21st century skills.” The top skills requested were oral and written communication, collaboration, problem-solving, communication skills, social intelligence, and self-direction. Shafer et al., (2023) similarly found written communication to be the most frequently requested skill (67%) for bachelor-level geoscientists. Mankki (2023), summarizing job ads for teacher educator jobs, found that the most frequently listed personal qualities were interpersonal skills and teamwork, sensitivity to cultural diversity and intercultural understanding, professionalism, and leadership skills.

Burning Glass (2018, p. 14) conducted an analysis of job ads, dividing skills sought into three categories: technical skills, software skills, and 142 baseline skills. The top baseline skill across all of 18 career areas was communication skills; teamwork and collaboration were rated in the top five across 14 areas; organizational skills were rated in the top 10 across 17 of the 18 areas, and problem-solving was especially highly rated in customer support and engineering.

Skills in higher education

A recent *Forbes* magazine article declared that “The soft skills debate is over” (Flynn, 2023). Flynn pointed out that since the publication of the influential Secretary’s Committee on Achieving Necessary Skills (SCANS) report in 1991, the importance of a foundation of basic skills (reading, writing, mathematics, speaking, and listening), thinking skills (creativity, decision-making, problem-solving, mind’s eye, knowing how to learn, and reasoning), and personal qualities (responsibility, self-esteem, sociability, self-management, and integrity) has been well-documented and yet employers lament that college graduates do not possess the critical soft skills or durable skills that are increasingly valued (Wilkie, 2019).

ETS has conducted several studies over the past 20 years that have asked college administrators and faculty members what qualities they believe are important for students entering higher education and what qualities are important to develop during higher education. These studies have used a variety of approaches—interviews, focus groups and, surveys. Among the most frequently nominated attributes are ones related to perseverance (grit, resilience, drive, work ethic), professionalism (organization, time-management, self-discipline, dependability, reliability), motivation, and passion for the field. Oswald et al., (2004), analyzing college mission statements, identified 12 dimensions of college performance, clustering into intellectual behaviors (knowledge, learning, and mastery of general principles; continuous learning and intellectual interest and curiosity; artistic appreciation and curiosity), interpersonal behaviors (multicultural tolerance and appreciation; interpersonal skills; social responsibility, citizenship, and involvement), and intrapersonal behaviors (physical and psychological health; career orientation; adaptability and life skills; perseverance; ethics and integrity).

The National Academies of Science, Engineering, and Medicine (Hilton & Herman, 2017) sought to identify critical interpersonal and intrapersonal skills based on evidence that these were related to postsecondary persistence and success and that they could be enhanced through intervention. They identified eight: conscientiousness behaviors, sense of belonging, academic self-efficacy, growth mindset, utility goals and values, intrinsic goals and interests, prosocial goals and values, and “positive future self.” *However, the report also noted that the measurement of these skills was poor quality: almost exclusively self-report and only rarely were the psychometric qualities of the assessments—reliability, validity, and fairness—mentioned. This indicates a clear opportunity for the future of assessments*.

Skills important in K-12

One of the first large-scale efforts to assess social emotional learning skills in K-12 was the California Office to Reform Education (CORE) project which received a waiver from No Child Left Behind (NCLB) legislation to do so. There are now many reports on findings and lessons learned (Krachman et al., 2016; Meyer et al., 2018; West et al., 2017) but what is important for this section is how the project went about identifying the key skills and what they concluded. Krachman et al., (2016) recounts an initial meeting in 2013 with participation from social-emotional learning experts and CORE district representatives. Expert staff suggested themes of “meaningful, measurable, and malleable.” District staff prioritized identifying at least one interpersonal and one intrapersonal factor. After a voting process, four competencies emerged—growth mindset, self-efficacy, self-management, and social awareness, with a fifth, collaborative problem-solving, almost included, but held up to await the results from PISA 2015’s findings on collaborative problem-solving. (The identification of these dimensions somewhat reflects the influence of the Collaborative for Academic, Social, and Emotional Learning (CASEL), which in turn reflects influences from developmental and social psychology.)

Meanwhile, several studies have found evidence for the importance of the Big 5 personality factors in education (Conscientiousness, Agreeableness, Extraversion, Emotional Stability, Openness; Mammadov, 2022) leading to the OECD adopting that model for its SSES; (Chernyshenko et al., 2018; OECD, 2021). The Big 5 is a personality psychology model that is widely used in industry and military entrance or personnel selection and classification testing. In the end, the two frameworks are not fundamentally that distinct, despite their disparate origins (Soto et al., 2022). OECD (2023) is planning a follow-up to this study, pursuing alternatives to the rating scale measures used in the first study.

"Portrait of a graduate" is a framework, adopted by several states and supported by the National Association of State Boards of Education (Norville, 2022), that allows states to "better define the skills and knowledge students should master before they graduate high school." It involves adopting a competency-based education approach (Patrick, 2021), and defining profiles by engaging with stakeholders to determine emphasis areas, such as communication and critical reasoning. For example, South Carolina's "Profile of a South Carolina Graduate Competency Framework" proposes 12: read critically, express ideas, investigate through inquiry, reason quantitatively, use sources, design solutions, learn independently, navigate conflict, lead teams, build networks, sustain wellness, and engage as a citizen. Other states have developed or are pursuing similar ideas.

Another approach to identifying the value of different skills is to consider OECD's PISA program's special topic areas. PISA is administered every three years, beginning in 2000, and tests 15-year-olds reading, mathematics, and science each cycle; 81 countries participated in the most recent (2022) survey. In addition, PISA adds a fourth *innovative domain assessment*, measuring a cross-curricular competency, which varies. The process for identifying the innovative domain assessment involves negotiating with participating countries, so the subject matter identified can be an indirect gauge of the popularity of a topic around the world. Since 2012, PISA's innovative domain assessments have been on problem-solving, financial literacy, collaborative problem-solving, global competence, and creativity.

OECD's (2019) *Skills for 2030: Conceptual Learning Framework* was based on inputs from an international group of stakeholders involved in OECD's *Future of Education and Skills 2030* project. The report defined skills as "the ability and capacity to carry out processes and to be able to use one's knowledge in a responsible way to achieve a goal" and "part of a holistic concept of competency, involving a mobilization of knowledge, skills, attitudes and values to meet complex demands" (p. 4). The group prioritized skills in 3 areas, cognitive and meta-cognitive skills (critical thinking, creative thinking, learning to learn, self-regulation); social and emotional skills (empathy, self-efficacy, responsibility, and collaboration), and practical and physical skills (using new information and communication technology devices). They also note that knowledge, and attitudes and values are intertwined and integral to developing knowledge and skills. They argued that cognitive skills are essential for solving complex problems and working with AI in complementary ways. Creativity is likely to remain viable, and higher-order skills such as problem-solving and critical thinking will remain important. Meta-cognitive skills are key to lifelong learning, which will become increasingly important with AI developments. Cultural understanding and dealing with uncertainty are also key to adapting to changes brought on by technology advances. Social and emotional skills are now recognized as essential and will remain so with demographic and societal changes; AI also is not likely to replace workers in occupations requiring social and emotional skills. Practical and physical skills, including ones involving the arts and the development of healthy habits and exercise routines, will continue to benefit individuals by supporting health and well-being.

The skills in demand are hard-to-measure skills

The previous subsections in this section have outlined a case for the assessment of various skills—learning to learn, creativity, communication, critical thinking, cultural understanding, curiosity, flexibility, resilience and others—for which there is no clear consensus yet on how they might best be measured. The National Academies of Science, Engineering, and Medicine (2017) report concluded that the measures that do exist, at least those in use, are poor—they are primarily rating scale self-reports accompanied by minimal or no basic information on their quality (reliability, validity, fairness). Stecher and Hamilton (2014) referred to the skills reviewed thus far as "hard-to-measure competencies." They concluded that "there is a need to develop a clear, comprehensive research agenda related to academic mind-sets, collaboration, oral communication, learning to learn, and other hard-to-measure 21st-century skills and competencies." (p. 71)

Traditional testing is not yet up to the task of measuring these skills routinely, but employers (historically) and admissions staff (increasingly) believe these skills are important and they will therefore evaluate these skills subjectively, through interviews, resumes, recommendations, or by self-reports. Yet, there have been significant developments over the past decade in methods for measuring hard-to-measure skills. For example, computer platforms have been designed to measure collaboration and collaborative problem-solving and using simulations (Hao et al., in press). Games and game-based assessments have been developed to assess a broad variety of skills (Landers & Sanchez, 2022) and personality (Landers et al., 2022) and are increasingly used in operational settings (Buckley et al., 2021). Situational judgment testing is now quite commonplace in the workforce (OPM, n.d.) and increasingly in educational settings (Wolcott et al., 2020) and can be gamified (Landers et al., 2022). We review these developments in Section 3.

Forecasting future skills demands

The nature of skills demanded in the workplace is changing due to technology—many skills valued today are likely to be automated soon; new skills that are not yet recognized will emerge. This change will affect education as well as the workforce. But how do we know what skills will be phased out and what the emerging skills might be? No one can reliably predict the future but it is safe to assume much of the future will be similar to the present. Thus, a useful starting point is to assume that the skills to be required in the school and workforce of the future will largely be those that we have just reviewed. However, here we will explore two additional methods, trends analyses of skills requirements in the workforce, and then task analyses of occupations to determine what jobs or parts of jobs might be vulnerable to technology displacement or complementarity.

Trends analyses

Several studies in Economics have examined workplace trends to determine the value of skills in the labor market. Using data from the U.S. Department of Labor's Dictionary of Occupational Titles, Autor et al., (2003) examined jobs from 1960 to 1998 and found that over that period technology was able to replace routine cognitive and manual tasks. Technology also placed new cognitive demands on workers, and what was valued in the workplace. Technology complemented activities involving non-routine work and interpersonal tasks, leading to a decline in manual work and routine cognitive work, but a rise in other kinds of work.

A similar phenomenon occurred due to communications technology (e.g., internet, social media) (2000-2015) placing new demands on social skills (Deming, 2017). Social skills allow for more efficient teamwork (according to Deming, workers trade tasks to exploit comparative advantage). Deming (2017) showed that from 2000–2012, the fastest-growing occupations were social ones, such as teachers, managers, nurses, and therapists. Non-social STEM occupations, such as engineers, drafters and surveyors, architects, and biological and physical scientists experienced negative growth. Social occupations grew by 12% as a share of all jobs in the U.S., and wages grew more rapidly. Weinberger (2014) found growth in employment and earnings for occupations requiring high levels of both cognitive and social skills compared to those requiring only one or the other.

Langer and Wiederhold (2023) examined data from German apprenticeship records that indicated the level of cognitive, social, digital, manual, management, and administrative skills training trainees received during apprenticeship. They found that a month of apprenticeship was worth two to three months of schooling with respect to higher wages, that returns were highest for digital, then social, then cognitive skills, and that apprenticeships that increased both cognitive and social skills provided the greatest returns, indicating skill complementarity, consistent with previous findings (Deming, 2017; Deming & Kahn, 2018; Weinberger, 2014).

Predictive AI effects on future jobs

Much has been written about the effects of AI on jobs in the future. For convenience, we will divide these works into two phases, the first focusing on machine learning, sometimes referred to as predictive AI; and the second on large language models and generative AI. Agrawal et al.'s (2022) book, on the disruptive economics of AI, was written from the standpoint of the value of predictive AI. In it, they argued that AI would assume the roles of prediction and judgment, which goes considerably beyond the automation of routine cognitive tasks observed by Autor et al., (2001), to complex cognitive tasks that had been assumed to require uniquely human capabilities. Agrawal et al., focused on the potential synergistic advantages, and on how work could be, and likely will be reorganized to achieve those advantages, but another implication is that even high-skill occupations will be exposed to AI, rather than the low-skilled and middle-skilled jobs exposed in earlier technology advances.

One of the first studies to put numbers to the incursion was by Frey and Osborne (2017), who studied jobs' susceptibility to computerization through an expert rating study. Experts rated 70 occupations sampled from the O*NET database judging which could be fully automated. Based on this they identified nine O*NET skills that were not susceptible to automation by computers or robots and matched these back to a larger list of 702 occupations to make estimates of the susceptibility of the labor force and certain occupations to automation, concluding that 47% of employment was at risk. The 9 non-susceptible skills were assisting and caring for others, persuasion, negotiation, social perceptiveness, fine arts, originality, manual dexterity, finger dexterity, and cramped workspace.

A more recent expert rating study (Lassébie and Quintini, 2022) similarly used an expert ratings approach, but experts rated the automatability of skills and abilities instead of occupations allowing a more precise estimate of the impact of AI and automation on jobs. Among the O*NET skills least susceptible to automatability were management of personnel resources, complex problem-solving, negotiation, social perceptiveness, assisting and caring for others, technology design, management of material resources, active learning, service orientation, repairing, originality, persuasion, and active listening (on the opposite end were number facility, memorization, wrist-finger speed, selective attention, and static strength). The results largely agreed but Lassébie and Quintini (2022) added complex problem-solving and active listening, and dropped fine arts, working in cramped spaces, finger dexterity and manual dexterity. There also was disagreement on several skills, including consulting and advising others (AI can perform in some contexts), selling or influencing others (recommender systems perform well), instructing (some instruction activities can be performed well by AI), management of one's own time and the time of others (dynamic scheduling technology is effective but perhaps limited in applicability), oral and written expression (rapid progress in NLP was already showing strong performance in this area, even when this study was conducted), scheduling work and activities (AI task planning is well-developed), and visual abilities (AI vision has advanced considerably in the past 10 years, especially during Covid).

OECD (2023) conducted a larger study to put the findings from this and other studies into the larger context of labor markets and the employment outlook. They concluded that AI is likely to have a significant impact on the labor market but with considerable uncertainty about what the impact would be and what suitable policy actions would be to promote trustworthy use. There are indications that AI creates new tasks and jobs for high-skilled workers with the right competencies and that AI can reduce tedious tasks and increase engagement and safety. But this might also leave a more intense, high-paced work environment. AI work management can increase perceived fairness but risk privacy and introduce or perpetuate biases. A key policy implication is that there is a growing need for education and training to ensure that workers have the skills to use the new technology.

Generative AI effects on future jobs

The second phase followed the release of OpenAI's ChatGPT in November 2022, and GPT-4 in March 2023, focusing on large language models, referred to as generative AI. There are other systems including Google's Gemini, Anthropic's Claude, and LLaMa. There are also text-to-image generative AI systems including Stable Diffusion, Midjourney, and DALL-E. The AI expert community was generally aware of and using developments in the underlying technology prior to the commercial releases (Lassébie and Quintini, 2022), but as noted by Cotra (2023) the capabilities of ChatGPT caught even the expert AI community by surprise, some suggesting that forecasts were that the capability level would not be achieved for another 10 or 20 years or even farther in the future.

One study on the possible impacts of ChatGPT and associated technologies, which they suggested may have the potential to become general-purpose technologies (see Bresnahan, 2010, for a review of general-purpose technologies, such as steam and electricity) was conducted by OpenAI (Eloundou et al., 2023). Their approach was to determine "exposure" of tasks and jobs to large language models (LLMs) without distinguishing between labor-augmenting and labor-displacing effects. They used both human raters (annotators) and GPT-4 to apply a rubric measuring exposure of tasks to LLMs, based primarily on O*NET. They concluded that 19% of jobs have at least 50% of their tasks exposed considering current LLM capabilities and associated tools, but by adding other generative models and complementary technologies, 49% of workers could have half or more of their tasks exposed. Regarding skills, roles reliant on science and critical thinking show a negative correlation with exposure (i.e., less susceptible to LLMs) but programming and writing skills are positively associated with LLM exposure (i.e., highly susceptible to LLMs).

The Eloundou et al., (2023) analysis did not result in conclusions dramatically different from those in previous studies, such as those reviewed here. Eloundou et al., (2023) computed correlations between their estimates of exposure and those obtained in other studies and found them generally to be positive and statistically significant. They did not include the Lassebie and Quintini (2022) study in their analysis, however.

It is possible to speculate also on the emergence of new jobs or new skills. For example, an important emerging skill is likely to be working with assistive technologies (e.g., LLMs). Already, digital literacy is a skill that appears in some employer surveys and in other contexts. However, digital literacy is such a broad concept—covering "...everything from reading on a Kindle to gauging the validity of a website or creating and sharing YouTube videos" (Loewus, 2016)—that interpreting statements about its importance is difficult. However, using ChatGPT and other generative AI technologies is already a valued skill and as the technology evolves, continuing to use generative AI technologies is likely to remain an important skill. The concept of personal digital assistants is not likely to disappear. The Computing Community Consortium and Association for the Advancement of Artificial Intelligence's (Gil & Selman, 2019) 20-year roadmap envisions a future world of personal assistants.

AI prompt engineer has also received considerable attention and is widely touted as a new, important occupation involving a set of new skills, the top-rated job of the future. However, Acar (2023) disagreed, arguing that future versions of generative AI systems will become more intuitive and less dependent on careful prompt crafting; in fact, he argued that AI models like GPT-4 are themselves very good at prompt engineering and they will likely get increasingly better. In addition, prompt engineering is LLM-specific, limiting its usefulness. Instead, he suggested that problem formulation per se is likely to emerge as an important skill, "the ability to identify, analyze, and delineate problems," along with the ability to interpret and critique the results from LLMs and then possibly reframe and execute again if needed. Acar suggested that four key components of problem formulation are problem diagnosis, decomposition, reframing, and constraint design and that these skills will become key to effective collaboration with AI systems.

Conclusions for Section 2: Skills for the future

For most of the past century, efforts and advances in assessment have primarily related to assessment of curricular skills—math, reading and science—that is, the skills targeted by the traditional K-12 educational curricula. Consequently, these skills have been the focus of large-scale domestic and international assessments, which are designed to monitor states' and nations' educational systems. Those skills are and will remain important, but for the past 20 years or so there has been a growing appreciation for the importance of other kinds of skills, which are now recognized as at least equally important—collaboration, problem-solving, critical thinking, creativity, curiosity and work ethic. Sometimes, particularly in the workforce sector, these are referred to as durable skills, indicating their generalizability and usefulness across all kinds of education, training, and workforce tasks and contexts. These skills are more challenging to measure and so can be referred to as hard-to-measure skills. With advances in technology and AI, there likely will continue to be changes in what skills are most valuable. Already we see that AI can perform language tasks, artistic creation, and coding tasks at the level of advanced college graduates and beyond.

This situation presents a challenge and an opportunity for assessment. The challenge is that simple self-assessment ratings that we rely on today are not sufficient for the task of providing useful information about the skills that will become increasingly important. The opportunity is that new innovative assessment methods can be developed so that we will be able to assess hard-to-measure constructs with the same level of sophistication that we are able to measure mathematics, reading and science today.

SECTION 03.

Innovative measures:

new approaches for assessing hard-to-measure skills

In this section, we review the methods that are currently used to measure hard-to-measure skills. For some hard-to-measure skills tests have been developed. Many others rely predominantly on self-reports and others' reports. We discuss their limitations and ways they can be improved. We focus on efforts to develop performance measures of hard-to-measure skills, which include situational judgment tests (SJTs), games, simulations, and interactive tasks. We also discuss approaches for measuring process data which allow us to make inferences about test-takers' skills through the actions, response times, and conversations they have while solving problems or interacting.

Setting the stage

The previous section discussed the skills that might become increasingly important in the future. The purpose of this section is to discuss methods for measuring those skills. Separating skills from how we measure them is not straightforward. Skill types are often confounded with measurement methods. For example, technical skills, such as coding, and traditional curricular skills, such as mathematics, reading comprehension, and writing, tend to be measured with tests, using multiple-choice, constructed response, and some more innovative answer formats. But there are no tests for many soft skills, which instead rely on subjective interviews or rating scale self- or other-report assessments. This is problematic because subjective measures may be perceived as less informative than tests. The LinkedIn Talent Solutions (2019) Global Talent Trends report found that 91% of talent managers believed that soft skills were important to the future of recruiting and 92% that soft skills matter as much or more than hard skills, but that 57% struggle to assess soft skills accurately. Managers reported using subjective judgments such as social cues in interviews, and that the dominant methods for assessing soft skills were behavioral questions (75%), reading body language (70%), and situational questions (58%), all subjective methods.

Table 9 presents a list of testing methods or item types from a variety of perspectives. Scalise and Gifford (2006), building on Bennett (1993), proposed a taxonomy of item types for computer-based assessments focusing on academic subjects; RAND's (2020) assessment compendium organized K-12 educational assessments by testing type to facilitate look up by educational practitioners, expanded to include social and emotional constructs. IMS Global's question test interoperability (QTI) standards are designed to support all digital assessments; interaction types, which are listed here because they are pertinent to the content of this section of the paper, represent a part of a broader set of interoperability standards. The Institute of Medicine's (2015) and the Office of Personnel Management's list of assessment methods presented methods used for clinical assessment and organizational recruitment and selection, respectively, and therefore use the language used by practitioners in those fields.

Several issues become apparent when reviewing the lists presented in Table 9. There are many, diverse approaches to assessment. Construct and method are often confounded—the Institute of Medicine (2015) and OPM both refer to personality tests, which is both a set of constructs but also a methodology (rating scales); they also refer to cognitive tests and cognitive ability, which similarly is both a testing method and construct. RAND's list reflects one search dimension, another one being construct (interpersonal, intrapersonal, cognitive). Scalise and Gifford (2006) as well as the IMS Global (2022) QTI standards, although in different ways and for different purposes, are both attempts to separate construct and method by focusing on the specific ways responses can be elicited from individuals to generate data enabling construct

TABLE 9.

VARIOUS PERSPECTIVE ON TESTING METHODS AND ITEM TYPES

Computer-Based Assessment of Cognitive Item Types[1]

- Multiple Choice
- Selection/Identification
- Reordering/Rearrangement
- Substitution/Correction
- Completion
- Construction
- Presentation

Psychological Assessment Measures and Methods[2]

- Screening instruments
- Checklists
- Questionnaires
- Memory tests
- Interview observations
- Personality tests
- Interviews
- Observations
- Cognitive tests
- Rating scales

Academic, Social, and Emotional Learning[3]

- Paper/Pencil Digital
- Oral
- Selected Response
- Performance Task

Employability Tests[4]

- Accomplishment Records
- Assessment Centers
- Biodata
- Cognitive Ability
- Emotional Intelligence
- Integrity/Honesty Tests
- Personality Tests
- Reference Checking
- Situational Judgment Test
- Structured Interviews
- Training and Experience
- Work Samples

QTI Standards[5] (Interaction Types)

- Choice
- Text Entry
- Extended Text
- Gap Match
- Hot Spot
- Hot Text
- Inline Choice
- Match
- Order (with Graphics)
- Associate (with Graphics)
- Media
- Position Object
- Select Point
- Slider
- Upload
- Drawing
- Custom
- End Attempt

SOURCES: [1]Scalise & Gifford (2006);[2] Institute of Medicine (2015); [3]RAND (2020); [4]OPM (n.d.); [5]IMS Global (2022).

interpretations, regardless of construct. Construct-method separation is a feature of the evidence-centered design framework (Mislevy et al., 2003).

It is possible, in principle, to measure the same skill in multiple ways. The multitrait-multimethod approach (Campbell & Fiske, 1959) and modeling framework (Kyriazos, 2018) is designed to reflect just this. For example, the construct curiosity could be measured with a self-rating (e.g., “I like to know how things work” Strongly disagree/disagree/neither agree nor disagree/agree/strongly agree), a teacher rating (e.g., “X likes to know how things work. True or False?”), a record from a computer log file indicating the number of times the student explored options; a performance test (e.g., the number of doors opened/paths marched down in a computer adventure game), a SJT (e.g., “You have an assignment to complete a five page research paper on one topic, but during your research you run across an exciting new approach to solve problems on an unrelated topic. What do you do?”), or a behavioral interview (“Tell me about a time when your curiosity led you to discover something interesting or useful"). A prediction for the future of assessments is that there will be increased attempts to use the methods expressed in Table 9, many of which originally were designed to measure technical and cognitive academic constructs, to measure some of the hard-to-measure constructs discussed in Section 2.

An example of how this might be done is found in Roll and Barhak-Rabinowitz's (2023) proposed approach to measure a hard-to-measure skill, self-regulated learning (SRL), on the PISA 2025 Learning in a Digital World (LDW) assessment. SRL is a composite construct, reflecting cognitive and meta-cognitive processes, affective regulation, and motivation (Molenaar et al., 2023). The most common approach for measuring self-regulation is the self-report questionnaire, which is prone to *reference bias* (Lira et al., 2022), the "systematic error arising from differences in the implicit standards by which individuals evaluate behavior" (from the Abstract). Roll and Barhak-Rabinowitz (2023) proposed instead a framework for measuring SRL through learners' (or test-takers') actions on a learning task that provides the *affordances* of allowing learners to experiment (e.g., through interactive simulations), to receive feedback (automatically or by pressing a button), and to seek information (e.g., watch tutorials, ask for hints, view worked examples). Roll and Barhak-Rabinowitz mapped these to the affordances provided in the PISA LDW assessment to identify how the assessment of SRL skills could be accomplished.

The Roll and Barhak-Rabinowitz (2023) paper was part of a collection of papers (Foster & Piacentini, 2023) related to using innovative assessments to measure complex skills. The executive summary for the collection highlights the importance of: (a) measuring what matters rather than what is easy, (b) assessments set in authentic contexts and involving learning, (c) innovation across all phases of assessment design, (d) digital technologies expanding what can be measured but needing better measurement models, and (e) the importance of validation.

All of these are high-priority research topics for the future of assessments, although with different emphases for different applications. We expect that research attention will increasingly be given to the important but hard-to-measure skills. Authentic learning contexts is not a new assessment topic (Erwin & Sebrell, 2003; Frensch & Funke, 1995), but technology developments may increase the usefulness of authentic assessments. Learning also is not a new topic in assessment; there is a substantial older literature on dynamic assessments (Grigorenko & Sternberg, 1998). But there are also promising newer measurement approaches that are likely to be pursued in the future in psychology (Bolsinova et al., 2022; Deonovic et al., 2018; Yeung, 2019) and economics (Heckman & Zhou, 2021).

For the remainder of this section, we review findings from the predominant methods for assessing hard-to-measure skills. We organized this by ratings and rankings, SJTs, performance measures and multimodal measures.

Ratings and related methods

Rating methods are ones in which raters rate attributes of themselves or others, typically on a rating scale, such as a Likert scale, although there are variants such as checklists. Rating methods, particularly self-ratings, are widely used, adaptable to just about any psychological or educational construct imaginable, and relatively inexpensive to develop, administer, score, and report on, which at least partly explains their popularity. Psychometric models for rating scale methods and nomological networks of constructs based on rating scale methods are well-developed. *It is likely that the world will continue to rely on rating scales for many skills and constructs for the foreseeable future*. Whole areas of psychological constructs such as personality (John & Srivastava, 1999), interests (Su et al., 2019), and others (Kyllonen, 2016) have been mapped based entirely on the rating scale methodology.

There are problems with rating scale measures, such as biases and other limitations of self- and other reports (Hoyt & Kerns, 1999; Salgado & Moscoso, 2019). Self-reports are subject to response style bias (van de Vijver & He, 2016), reference bias (Lira et al., 2022), social desirability bias (Paulhus, 2002), and faking (Geiger et al., 2021), which is particularly problematic when used in high-stakes assessment situations (Niessen & Meijer, 2017).

Ratings by others, or informant ratings, mitigate, at least to some extent, social desirability and faking (informants of course, can provide socially desirable or faked ratings on behalf of the target and may if sufficiently incentivized to do so). Ratings by others have also been found to be better predictors of future behavior compared to self-ratings (Connelly & Ones, 2010; Oh et al., 2011; Poropat, 2014). Letters of recommendation are only weakly correlated with higher education performance but still predict outcomes such as degree attainment (Kuncel et al., 2014).

Ranking methods including **forced-choice** approaches ask respondents to rank rather than rate attributes of themselves to mitigate socially desirable responding and response style bias. This can be particularly useful in high-stakes settings and consequently there is considerable interest in using such approaches in admissions. Developments in scoring forced-choice methods have led to an increase in the reliability of the method (Fu et al., in press), which already has been shown to have higher predictions of outcomes compared to rating scale methods (Cao et al., 2015; Salgado & Tauriz, 2014).

Anchoring methods are another approach to reducing response biases in rating scale methods and primarily response style bias. Anchoring vignettes (King & Wand, 2004) ask respondents to rate themselves and a hypothetical other and then scale the self-ratings to the scale established by rating the common (across respondents) hypothetical others. This approach has been shown to reduce country-level differences in response styles increasing construct comparability across countries (Kyllonen & Bertling, 2013). Ludlow et al., (2022) showed how a related technique can be used to measure the hard-to-measure skill, *life purpose*. For ratings by others (e.g., teachers, peers, supervisors), behaviorally anchored rating scales (BARS) also use anchors as rating aids and are widely used in organizational program evaluation studies (Kell et al., 2017; Klieger et al., 2018); BARS are not typically used for self-assessments but for ratings by others.

Situational judgment tests

SJTs present a situation description and ask respondents to indicate how they would respond or what the best response would be in the situation. Figure 3 provides an example. SJTs are a popular method for measuring hard-to-measure constructs, particularly interpersonal skills (Christian et al., 2010). SJTs are a flexible method and can involve written materials or videos, and typically ask for response options to be ranked or rated. SJTs are widely used in organizational contexts (U.S. Office of Personnel Management, n.d.) for employee screening and sometimes for training (Cox et al., 2017).

FIGURE 3.

EXAMPLE SITUATIONAL JUDGMENT TEST (SJT) ITEM FROM ZU & KYLLONEN (2020)

You're one of the managers for a large volunteer agency. In a discussion about how to find new volunteers, you bring up what you think is a great new idea. But the other managers tell you that the idea is "off base" and not workable. How would you handle this situation?

A. Drop your idea because the group is probably right.

B. Point out several good reasons why your idea might work.

C. Drop your idea for now but tell it to your boss later.

D. Tell the other managers that lots of people don't recognize great ideas at first.

SJTs have also been used in educational contexts (Sternberg et al., 2000; MacCann & Roberts, 2008). The College Board experimented with SJT undergraduate admissions tests (Schmitt et al., 2009) and SJTs have been used in business school (Hedlund et al., 2006) and dental school admissions (Buyse & Lievens, 2011). The American Association of Medical Colleges (AAMC, n.d.) currently offers a 75 min SJT called PREview for medical school admissions to measure nine professional competencies including Interpersonal Skills, Cultural Awareness, Cultural Humility, Empathy and Compassion, Teamwork and Collaboration, Ethical Responsibility to Self and Others, Resilience and Adaptability, Reliability and Dependability, and Commitment to Learning and Growth. Acuity Insights (n.d.) offers a competing test, Casper, which is a 90-minute open-ended SJT measuring 10 aspects of social intelligence and professionalism with 14 scenarios (eight typed, six video) and publishes a manual (Acuity Insights, 2023).

SJTs have proven to be valuable for their flexibility in use and in their appropriateness for assessing hard-to-measure skills, and therefore are likely to remain popular as assessment methods in the future. However, there are challenges. SJTs tend to be less reliable than rating scale measures per minute of testing time and therefore tend to take longer or take more testing time to get a reliable score. Although Casper is a 90-minute test, it measures a single factor, not the 10 aspects. This is a general finding with SJTs. For example, while Oswald et al., (2004) developed 12 higher education competencies (e.g., leadership, artistic), the SJT they developed measured only a single dimension (Schmitt et al., 2009). Given that SJT research and promotional material indicate a desire to measure multiple dimensions with SJTs, a future research challenge for SJTs generally will be to measure multiple dimensions reliably in a reasonable amount of time.

Performance measures

Performance measures for some of the key skills identified in Section 2 are reasonably well established, such as ones for critical thinking (Liu et al., 2016) and creativity (Weiss et al., 2021). The future of assessments will include such measures for applications and will see further incremental development. Another significant development in the measurement of these skills will be related to how these skills increase over time. Koedinger et al., (2023) were able to discover a regularity in learning rates once initial knowledge levels are taken into account based on data collected from over a million observations from lessons of math, science, and language distributed with intelligent tutoring systems. Duolingo's Birdbrain system adapts language instruction to deliver an exercise at an optimal difficulty level for engagement and learning (Bicknell et al., 2023). It does this by predicting student performance based on ability and item difficulty using item response theory (IRT) then updating student ability level following exercise performance, in a manner similar to how it is done in adaptive testing. The merging of traditional assessment of skills with concepts related to updating skills as is done in adaptive testing and in adaptive instruction and hybrid approaches is likely to be increasingly important in the future of assessments as is the use of hybrid IRT modeling in adaptive instruction applications (Scalise et al., 2023; Yeung, 2019), particularly as electronic instruction and record keeping becomes even more ubiquitous.

We believe that another significant push will be in the development of performance measures for constructs and skills that are now largely measured by checklists and rating scales—consider the typical employment interview assessing various candidate qualities such as self-motivation, originality, and time-management, through questions and a rating rubric. The kinds of constructs we believe will be amenable to performance measurement include teamwork, collaboration, leadership, self-management and self-regulation, emotional management, work ethic, flexibility, cultural sensitivity and other soft skills or durable skills listed in Table 8. Attempts to measure such soft skills with performance measures has a long tradition, sometimes under the name of objective personality tests (Cattell & Warburton, 1967; Ortner et al., 2006). Alan et al.'s (2019) grit game and Segal's (2012) use of persistence in the coding speed test as a measure of intrinsic motivation are examples of performance measures of personality traits. Charness et al., (2018) provided a list of real-effort tasks used in behavioral economics studies that can be understood to be measures of the traits of persistence, self-management, or conscientiousness. Kyllonen and Kell (2018) summarized much of this literature, dividing into the categories of low-stakes cognitive tests (Segal, 2012), objective personality tests (Ortner & Proyer, 2015), economic preference tasks (Falk et al., 2016), confidence judgments (Stankov et al., 2013), survey behavior (Soland & Kuhfeld 2018), item position effects (Weirich et al., 2017), and effort inferred from response time (Wise, 2014). All of these can be understood as attempts to measure soft skills with performance tasks rather than ratings.

Collaborative problem-solving is an example of a performance measure of a soft skill, or set of soft skills such as communication, teamwork, and collaboration. ETS has developed an assessment platform (ETS Platform for Collaborative Assessment and Learning, EPCAL; Hao et al., 2017), and a set of tasks including negotiation (Martin-Raugh et al., 2020), letters-to-numbers problem-solving, hidden profile decision-making (Kyllonen et al., 2021), and others, which measure both team performance and individual collaborative skills (Hao et al., 2019). Given the importance of social skills in the future workforce (Deming, 2017), it is likely that these kinds of assessment projects will grow in importance in the future.

Life data (L-data)

Cattell (1965) proposed the use of L-data (life data) defined as "behaviour in the actual, everyday life situations" to serve as the basis for inferences about targets' skills. The widespread availability of all kinds of records—administrative, social media, cell phone and web sites—makes the collection of such L-data much easier than it was when Cattell proposed using it. This kind of data is substantially different from what is obtained from responses on questionnaires or SJTs. These are the traces and indicators individuals leave behind from which inferences about individual attributes might be drawn. Studies have used administrative records to create composites that reflect students' academic and non-academic skills and behavior to help predict graduation and to evaluate teacher effects (Jackson, 2018; Kautz & Zanoni, 2014; Novarese & di Giovanni, 2013). Social media and other behavioral traces have been treated as reflections of personality (Gosling et al., 2011; Kosinski et al., 2014; Youyou et al., 2015). For example, Gosling et al., (2002) measured personality based on the appearance and contents found in dorm rooms. Background and experiences can also be measured with traditional surveys, structured resumes, or biodata (Mumford et al., 2013), ambulatory assessments (Trull & Ebner-Priemer, 2013), and mobile sensing data from phones, wearables, and beacons (Mattingly et al., 2019; Mirjafari et al., 2019). The future of assessments may involve increasingly the incorporation of diverse kinds of data that might be used to draw inferences about students' and workers' knowledge, skills, abilities, behaviors, values, and attitudes. However, privacy concerns will have to be addressed, particularly in light of the 2024 EU AI Act (see footnote 5), anticipated upcoming regulation in the U.S. and elsewhere, and general AI ethics statements being issued by many organizations (Abrams, 2024; Blackman & Ammanath, 2022).

Game-based approaches

Game-based assessments can be defined as "an assessment method in which [examinees] are players participating in a core gameplay loop while trait information is inferred" (Landers & Sanchez, 2022, p. 1). Landers and Sanchez (2003) also defined the related concepts of *gamified assessments*, as ones in which game mechanics or game concepts are applied to existing, traditional assessments and *gamefully designed* assessments in which test developers use game concepts in designing new tests. There are several reasons why a game-based or gamified assessment might be used. For high-stakes purposes, for example in employee selection, which is the application Landers and Sanchez (2023) focused on, the assessment can serve not only to measure an applicant's skills, but also to serve as possibly a realistic job preview, or as a recruiting device in providing "compelling experiences" to candidates (p. 21). Game-based assessments also might be uniquely capable of measuring certain hard-to-measure skills, such as curiosity and social preferences (Tang & Kirman, 2023). A third reason is to increase test-taker engagement. Test-takers in high-stakes testing situations already have sufficient incentives to be engaged in test-taking; but this is not true in settings that are low-stakes to the test-taker, such as in school accountability testing, domestic and international large-scale assessments, and research settings. In these settings, a lack of motivation can affect scores (Liu et al., 2012), and a gamified version of the test might lead to increased engagement and motivation and consequently better reflect test-takers' skills. Buckley et al., (2021) reviewed game-based and gamified assessments in education including SimCityEDU: Pollution Challenge (Mislevy, 2014), ACTNext's Crisis in Space (Chopade et al., 2019), and Imbellus's Project Education Ecosystem Placement (PEEP).

Multimodal measures or process data

Multimodal measures can be defined as ones using physiological data (e.g., EEG, heart rate), behavioral data recorded in log files (e.g., conversations, chats, keystrokes, eye tracking), and poses and facial expressions captured in audio and video recordings analyzed by human raters or automatically (Molenaar et al., 2023; Slavich et al., 2019). Such data are beginning to be used in assessment applications and their use will likely increase (Martin-Raugh et al., 2023). For example, an ETS project by Chen et al., (2014) analyzed public speaking skills by giving 17 speakers four speaking tasks and then capturing performance with audio, video, and 3D capturing devices. They extracted features using natural language processing (NLP) methods, speech processing, and multimodal sensing (using Microsoft's Kinect [Zhang, 2012]) to capture both speech and non-verbal communication, including hand gestures and head orientation. Chen et al., (2014) developed a scoring model for the data extracted and found that the resulting scores correlated with human holistic ratings of public speaking performances. In two other ETS projects, Martin-Raugh et al. (2020) and Jiang et al. (2023) analyzed conversations during negotiation and collaborative problem-solving to gain insights into the processing differences between successful and less successful collaborations. They used NLP methods to classify conversational turns to a rubric, which comprised categories such as greetings, sharing information, acknowledging contributions, and negotiating and found correlations between the nature of conversations conducted by individuals and teams and task success. Much of the history of assessment has been based on limited interactions and a very limited data stream between the test and test-taker. *Multimodal assessment opens the door to much richer forms of expression from which inferences about what test-takers know and can do can be made.*

Conclusions for Section 3: Innovative measures

The predominant method for measuring the kinds of skills reviewed in Section 2 are rating scale methods. That is because the skills highlighted are hard-to-measure skills and rating scales are a general and flexible approach for measuring just about any skill for which a definition can be articulated. There are ways to improve over self-report ratings. Other reports are less susceptible to biases associated with self-reports, such as reference bias (Lira et al., 2022), but they have their own limitations, such as halo effects (Cooper, 1981). *Forced-choice measures also reduce biases associated with self-reports and they are therefore generally preferred over self-reports. SJTs also are a flexible measurement method that can be applied to many hard-to-measure skills*. They have a conceptual advantage over self-reports in that they can measure knowledge of correct or proper or useful actions rather than simply general assessments of typical behaviors. The future of assessments will likely involve a move away from self-reports to these other forms of measurement, more routinely.

However, we believe the more significant movement in the future of assessments will involve the development and adoption of performance-based measures, such as games and interactive tasks, such as an actual negotiation session or a collaborative problem-solving task, to measure the important skills of the future. Performance measurement of personality has been a long-sought goal in the field (Ortner & Proyer, 2018) and some progress has been made (Kyllonen & Kell, 2018). Performance measures in principle have significant advantages over ratings: performance measures are not susceptible to ratingratings biases (e.g., halo, reference bias, response style, social desirability) and can be objective samples of behavior rather than subjective evaluations of behavior. (However, performance tasks that require observer ratings may still be subject to rating biases, such as halo/horn [Noor et al., 2023], severity/leniency [Cheng et al., 2017], and drift [McLaughlin et al., 2009].) Because performance measures are not yet well developed for many important constructs, these constructs therefore continue to rely on subjective ratings for their measurement. We believe that performance measures will be supplemented by test-less measures, involving process analysis and data mining, which can be used to draw inferences about users' or students' or employees' skill levels (Baker & Yacef, 2009). There are good examples in varied domains ranging from social and emotional learning (Jackson, 2018; Kautz & Zanoni, 2014) to academic performance (Waheed et al., 2020), and STEM job participation (Yeung & Yeung, 2019).

SECTION 04.

Operations breakthroughs:

AI and technology-enabled advances

In this section, we consider the phases of the test development cycle from initial consideration of test purposes, administration and administrative constraints through item development, test assembly, security, quality control, scoring, and test evaluation, including validity and fairness considerations. A major theme throughout this section is that technology advances, particularly in AI, are likely to have a significant effect on all operations and all phases of test development.

Setting the stage

Testing operations refers to all phases of test development associated with fielding a test form, given a construct (defined in Section 2) and a testing method (defined in Section 3). This includes designing around test purposes and administrative constraints, developing items, assembling test forms, reviewing tests, delivering and administering tests, scoring, reporting scores, evaluating tests, item banking, managing all aspects of security from administration to scoring, and conducting quality control over the entire process. These processes are the core of the standardized testing industry. Schmeiser and Welch (2006) provided a comprehensive overview of how these phases have traditionally been accomplished and some of the key issues; International Test Commission (2001, 2013, 2017) and Association of Test Publishers (2022) supplemented this work with recent considerations from technology-based assessments (TBAs). Schmeiser and Welch (2006) described an evolution from an art to a science of test development over the past 60 years. This evolution can be seen in various phases, for example, from informal rules-of-thumb in test assembly to more recent use of mixed-integer programming (Davey, 2023; van der Linden, 2005); or from human- to machine-scored essays (Shermis & Burstein, 2013).

We believe the future of assessments will continue and perhaps accelerate the transition from an art to a science of test development, using advances in technology, operations research methods, and both predictive and generative AI methods.

In this section, we review current cutting-edge advances in the phases of test development to provide a basis for where we might see significant breakthroughs. Some promising ideas within the operations sphere are that technology will enable advances in both the fairness of assessments and in the trustworthiness of the information obtainable through assessments. Fairness treatments have traditionally been most developed for the tail-end of assessment development after test data are collected, to evaluate statistically whether item responses can be interpreted the same way for different groups of test-takers (Millsap, 2011). Technology promises to bring fairness sophistication forward to the initial stages of item creation, a process that has historically depended on informal policies and checklists (ETS, 2014, 2022). Security, too, represents an area where technology may bring benefits in ensuring that assessment scores are directly indicative of a test-taker's skills on the intended target construct and not compromised by unknown influences. We address these issues in this section.

Administration & administrative constraints

Time

A general rule in assessment is that the more time available for testing, or, the more information obtained from an individual, the more reliable the assessment will be. And the more reliable the information, the more likely that inferences drawn from the measurement will be justified and useful, for example, useful for predicting future outcomes. With additional information, the construct signal emerges increasingly clearly over the backdrop of measurement error noise. This is true in all assessment domains—from target practice to weight measurement—all other things equal, the longer the test, in items or time, the better. The problem in most cases is that individuals do not want to sit through long tests and sponsoring parties do not want to administer or pay for them.

There are several strategies for addressing this problem. Tests can be made more efficient through traditional psychometric approaches, by considering time as well as the amount of information a test item is provided. This was a major motivating factor for adaptive testing, which promised 50% savings in testing time (van der Linden & Glas, 2010). Multidimensional adaptive testing (Segall, 1996), which uses performance on related tests (or scales) to update score estimates for the test (or scale) being taken, takes this idea one step further, promising an additional 33% time savings for the same amount of measurement precision. A future application of this idea would, subject to privacy restrictions, mine all available information sources about a person—social media, educational records, letters of recommendation, resumes, voluntarily submitted materials—simply as the starting basis for skills estimation, updated with new information from the testing or assessment session. This could result in additional time savings in the testing session.

Another strategy would be to make the test-taking experience more useful for the test-taker in benefits returned. For example, instruction is typically coupled with assessment, but a learner receiving instruction receives a direct benefit of increased skill that may be perceived by the learner as a justified use of his or her time, even though the instructional time is at least partially spent taking an assessment. Formative assessments, intelligent tutoring systems, and other forms of instruction mixed with assessment take advantage of this principle to coax learners into spending more time being assessed, in principle, enabling better measurements of skill.

Still, another strategy is to make the test-taking experience more compelling or enjoyable so that test-takers willingly allow themselves to spend more time being tested. Multimedia tests, game-based, gamified, and gamefully designed assessments (Landers & Sanchez, 2022) take advantage of the fact that many children and adults enjoy playing games, and do so voluntarily without compensation, and with no perceived extrinsic direct benefit. DARPA's (2004) DARWARS program was based on the idea that students would voluntarily spend hundreds of hours participating in training experiences thereby acquiring skills through playing multi-player games, in virtual worlds, with simulations, intelligent agents, and online communities (O'Neil et al., 2004). Bandwidth challenges may be present in some circumstances and some parts of the world, which can be a validity threat.

Anytime, anywhere, with security

Anytime-anywhere test delivery has long been anticipated to meet a market demand for more convenience and less cost for the test-taker. Over 25 years ago, Bennett (1998) already suggested that dedicated test centers "may be on the endangered list." The Covid pandemic accelerated that push and now anytime-anywhere testing is a reality for many large-scale tests, which has had significant positive benefits regarding cost, convenience, and accessibility. Test centers still exist and they are still popular. For some people in some circumstances, it is more convenient and less costly to take a test at a test center and that might be true for a long time. In many cases, employers, higher education, and associations still require in-person testing for a variety of reasons, including security, adding more elements to assessments, the candidate experience, and fairness of access to digital connectivity and high bandwidth. The future may be a mix of increased at-home and even potentially mobile convenience with the continued availability of test centers.

Security is more challenging with at-home or mobile testing when the testing purpose is high-stakes. Security issues are currently addressed in various ways (Choi et al., 2021; ETS, 2023b; see Security and Quality Control; Qian et al., 2018a, 2018b) and will require continued monitoring and progress. For low-stakes, formative testing, where security is not as critical, there are many advantages to mobile testing. Karay et al., (2020), for example, found that the convenience of taking a test on a mobile device without time restrictions led to no differences in scores for anytime-anywhere vs. testing center locations, but significant advantages for mobile testing in engagement: students spent more time on the test and were more likely to use the help of books and online resources, consequently earning higher acceptability evaluations from students.

New devices

Some tests are still administered in the paper-and-pencil format, but increasingly rarely: even the SAT has gone fully digital this year (College Board, 2023). High-stakes graduate and professional school tests were transformed to digital-based assessments in the 2000s; large-scale domestic and international assessments serving countries all over the world, including many developing countries, converted to digital-based assessments in the 2015-2020 time period, although exceptions remain (OECD's PISA-D, administered in developing economies, is administered with traditional test booklets). Initial digital-based conversions did not add much functionality to the paper-and-pencil format, other than allowing many more test forms, and enabling adaptiveness, but increasingly new capabilities were introduced, allowing videos, simulations, and interactivity. These trends will continue, and increasingly more engaging and immersive formats will be possible, mirroring (but lagging) trends in the entertainment sector (lagging due to lower returns on technology investment in education versus entertainment).

New technologies, such as Apple Vision Pro's mixed reality headset, or Azure's Kinect, greatly expand the possibilities for how the inputs for testing—instructions, item stimuli, item prompts—and the responses to test items—gestures, grasping, whole-body movements—can be realized, allowing the testing of new constructs in new ways. A challenge here is that the technology changes rapidly due to market fluctuations, and investing in a technology to produce new kinds of tests carries risks. Microsoft stopped manufacturing Kinect in 2017 (Lee, 2023) and the promising sociometric badge technology (Lederman et al., 2016) was discontinued some time ago.

Item development

Automatic item generation using generative AI and item models

Item development has traditionally relied on human experts (Lane et al., 2016, provides a review) and has thus been an expensive and time-consuming process. Automatic item generation (AIG) is an appealing alternative that can make the process more efficient and standardized. Early AIG attempts (Irvine & Kyllonen, 2002) focused on building comprehensive *item models* of target knowledge, skills, and abilities, which were measured by prototypical items and then generated many comparable variants by manipulating key components of models. Those early attempts were effective in generating high-quality items similar to their respective parents but with two crucial caveats: they were difficult to scale because each item needed its own model, and they were limited in the variability of texts that provided context for the target construct, resulting in items that looked similar and therefore did not provide as much information as independent items (Bejar et al., 2002). Generative AI is particularly well-suited for overcoming these caveats; it can generate a wide range of texts across multiple item types. Recent AIG approaches thus frequently utilize LLMs to generate contexts, stems, and options for many different item types (e.g., Attali et al., 2022; Bezirhan & von Davier, 2023; Chan et al., 2022; Gao et al., 2022; Stowe et al., 2022; Zu et al., 2023). A successful combination of careful item modeling and capable LLMs appears a highly promising approach towards implementing automation into item development.

Writing an item, just like any form of writing, involves a process consisting of multiple steps. AIG approaches introduced in the literature thus far have exclusively focused on the initial generation. In this light, a more precise label for the current AIG approaches (including the LLM-based ones) would be automatic item *drafting*. To fully realize the potential of automation in item development, the entire process needs to be designed to utilize automatically generated item drafts. Draft items need to be reviewed for their accuracy, appropriateness, and fairness, calibrated to estimate their difficulty and discrimination, and assembled into a delivery unit (e.g., a test form) before they reach test-takers. The item development processes at ETS and many other testing companies were designed decades ago to accommodate a stable number of manually written item drafts arriving at regular intervals; the legacy processes can present a significant bottleneck in utilizing a large number of automatically generated drafts to achieve efficiency and scale in item development. It is thus crucial to innovate the overall item development process along with the initial generation capability.

Difficulty modeling using LLMs

Item difficulty plays a critical role in assembling test forms and determining scale scores. The standard practice is to estimate item difficulty based on a large sample of test-taker responses (often 500 to 1000 responses, that is, test-takers, per item). An implicit assumption under that practice is that the number of new items (whose item difficulty is estimated) is much smaller than the number of available test-takers. If an effective AIG system can instantly produce many new items, this assumption does not hold anymore, and the standard practice to estimate item difficulty becomes a major barrier to using newly generated items. Predicting item difficulty presents an alternative solution. In the past, such prediction attempts were hampered by the need for specific models for each item type as well as the limited capacity of prediction algorithms.

On the other hand, LLMs can be fine-tuned to serve as a generic and highly flexible modeling framework that takes an item as input and produces a prediction of its difficulty. Zu and Choi (2023a, 2023b) showed that fine-tuning an open-source LLM for item difficulty prediction outperformed the previous state-of-the-art prediction method (Loukina et al., 2016) as well as human expert difficulty judgment by a substantial margin. ETS researchers have also developed methodologies to account for increased uncertainty in predicted item difficulty for downstream psychometric tasks (Lewis, 2001; Mislevy et al., 1993).

Contextualization and personalization

Large-scale international assessments such as OECD's PISA, PIAAC, or SSES are administered in multiple countries and languages throughout the world, and country performance is compared in league tables, meaning comparability is key. Tests are prepared from English (or French) source versions, in a two-step process: adaptation and translation (cApStAn & Halleux, 2019; Hambleton, 2002). In the adaptation step, bi-lingual speakers from countries indicate whether a concept is meaningful and will be interpreted in the same way in their culture as it is in the source country. Consider a few of the ITC (2017) guidelines to get a sense of what adaptation entails:

- Minimize the influence of any cultural and linguistic differences that are irrelevant to the intended uses of the test in the populations of interest.
- Ensure that the adaptation process considers linguistic, psychological, and cultural differences in the intended populations through the choice of experts with relevant expertise.
- Prepare administration materials and instructions to minimize any culture- and language-related problems that are caused by administration procedures and response modes that can affect the validity of the inferences drawn from the scores.

The kind of adaptation called for, which is routinely applied in all international testing, requires expertise in the languages, and cultures involved along with the test content and testing principles. *It is a costly but necessary process to ensure comparability of assessment results across language and culture groups*. This work is not confined to international large-scale assessment work. Such adaptation is also called for in cases where a test, typically an employment test, is used locally in different countries (e.g., ETS's former Workskills for Job Fit was administered in 18 languages), and even with language subgroups within the United States (e.g., ETS's effort in cases in CA K-12 on cultural adaptation for Spanish speaking ELL test-takers). An argument could be made that a similar kind of adaptation could be applied to cultural subgroups within the U.S. Certainly, the kinds of biases identified in cross-cultural assessment—construct, method, and item bias—are relevant to the appropriate interpretation of test scores even for subcultures within a language group (van de Vijver & Poortinga, 2005).

A similar kind of adaptation is done in business or advertising, where, for example, proposals or ads for potential clients in multiple countries representing multiple cultures are prepared. Andi Mann, Sageable CEO and pioneer in the field of AI Ops, suggested that this kind of adaptation will soon be done with AI, "recontextualizing content for different cultures." He gave examples of taking a business proposal or advertising brochure, and adapting it to align with the target cultural values, for example, making it more culturally appropriate and less offensive, changing from a formal to informal tone and reframing goals to be more culturally compatible. He suggested that automatic recontextualizing will soon "become as simple as resizing images" (Turchin, 2023). Lee et al., (2023) demonstrated the use of LLMs for personalized marketing.

A related idea stems from the personalized learning literature (Walkington & Bernacki, 2020), which is defined as "adapting an experience or interaction to be appropriate for a specific person given a certain set of individual characteristics/qualities." Personalized assessment, then, is assessment of personalized learning. (See further discussion in Section 5.).

LLMs to accomplish personalization and contextualization

Technology and AI enable personalization or contextualization (economically and at scale) in test content generation to a degree not possible before. A high percentage (78%) of respondents in the ETS Human Progress Study (2023a) agreed that "AI has the potential to enhance learning assessments by tailoring them specifically to each individual learner's needs." This kind of personalization can be done within or outside the context of automatic item generation.

When test forms are assembled from a set of pre-made items, adaptation of any kind becomes a major challenge. Efforts to incorporate personalization in assessment need to overcome this challenge with the online adaptation of content for individuals or with a large enough pool of diverse items to approximate such a real-time experience. Neither is feasible under the current test development process relying on pre-made content. *Automatic adaptation of assessment content can thus be a highly impactful innovation that can accelerate the movement toward personalized assessment*. LLMs are well-suited to the automatic adaptation task, for they have already been pretrained on a wide variety of texts. Moreover, they can encode texts into numeric vectors with semantic information, which suggests that the success of neural style transfer approaches in computer vision (Gatys et al., 2015) may be applicable to the adaptation problem in the text domain (Hu et al., 2017; Prabhumoye et al., 2018; Shen et al., 2017; Yang et al., 2018).

However, to properly implement such an approach, its potential consequences need to be considered carefully. In the ETS Human Progress Study (ETS, 2023a), 71% of respondents worried that "AI has the potential to negatively impact learning assessments due to unintentional biases and programming flaws within the system." LLMs inherit biases[6] incorporated in their pretraining samples and may reproduce such biases in their output. A naïve LLM-based adaptation approach may thus lead to further dissemination and reinforcement of existing biases. Therefore, it is crucial to build a robust mechanism to monitor and control adaptation output to ensure that the benefits of automated adaptation can be realized without reproducing preexisting biases. ETS's history of, and expertise in, adapting large-scale educational surveys for multiple cultures and languages as well as the work undertaken to mitigate these issues in AI models that generate content is an important advantage in successfully addressing this challenge.

Test assembly

Test assembly is the process of selecting items for a test form. Selection is driven by constraints, which typically are specified on test blueprints (Davey, 2023; Lane et al., 2016). This is done to ensure comparability between forms, adherence to construct definitions, capture of all aspects of the content domain that the test is intended to measure (avoidance of construct underrepresentation), and the minimization of construct-irrelevant features. The kinds of constraints that can be included is essentially endless but typically relate to test length, construct, content (e.g., primary, secondary, and specific), item type or format, stand-alone or one-in-a-set item, and cognitive level or depth of knowledge (Davey, 2023). Psychometric properties such as item difficulty, item discrimination (the degree to which test-takers who are high versus low on the trait being measured are likely to solve the item correctly), and expected time to complete the item, can also be included, as can all kinds of details related to item content, such as the number of mentions of boys versus girls on a test form, typically balanced so as to minimize construct irrelevant influences affecting item responses. Even formatting (e.g., no longer than seven pages, maximum of 50 lines per page) can be included as part of the assembly process (Diao & van der Linden, 2013).

Since Stocking and Swanson's (1993) initial operational demonstration and van der Linden's (2005) compendium, the advantages of automated assembly, that is, treating test assembly as a combinatorial optimization problem, have been clear (Davey, 2023). This is the same technology airlines and the military use to fill seats and retailers use to fill shelves. In combinatorial optimization an objective function is minimized, subject to a set of constraints. The test blueprint is specified as a set of constraints, and the objective function is chosen to achieve some goal for the test design, such as achieving a target average item difficulty or target test characteristic curve (e.g., to create a form that provides information for all proficiency levels, or one that provides the most information around a cutoff point) or test information function or both (Ali & van Rijn, 2016). The objective function can also be used to achieve content goals or security goals (e.g., minimizing item exposure); (Davey, 2023).

[6] Here, we use the term bias in its original meaning rather than as a technical term in Statistics.

The combinatorial optimization approach to test assembly is incredibly powerful and capable of serving as the basis for realizing significant advances in the quality of tests. The approach is only limited by the data available on items to use as constraints and for the objective function. Davey (2023) recalled an incident of using automated test assembly to create a form in the early days of the technology, where the reviewing test developers discovered a flaw, obvious to them, but not to the algorithm, of including too many items that had “water” as a theme. Davey stated that there was no formal content requirement dealing with water, so the algorithm was blind to its overrepresentation, which stood out to the human reviewers.

The reason many item features, such as “water,” are not included in item banks is a combination of not previously having had a reason for including it (something easy for human writers to spot) and not having the time and labor to code items for every possible feature that might be relevant. Now that automated assembly approaches are in more widespread use, and that such approaches can easily handle many item features, this seems to be an area ripe for new approaches to solving the “water” problem, the problem of classifying items by large numbers of features without having to code them by hand.

Security and quality control

For the high-stakes testing context—school admissions, scholarship provision, selection testing for jobs—the testing industry uniquely provides reliable, valid, and trustworthy information about students’ or applicants’ skills to the employer or educational institution seeking that information for its own decision-making purposes. This is a major component of the standardized testing value proposition. Other sources of information about a person’s skills—letters of recommendation, personal statements, resumes—provide limited useful information about candidates’ skills and are highly susceptible to compromise and bias. Personal statements do not predict grades or faculty ratings after test scores and prior grades are accounted for (Murphy –., 2009), perhaps because they reflect inputs from sources other than the candidate (e.g., friends, family members, professionals) as much as from the candidates themselves (Powers & Fowles, 1997). Reference letters are more predictive of outcomes than are personal statements (Kuncel et al., 2014), but they have their own problems: there is a strong bias against negative comments, the level of agreement between two raters tends to be small, letter writers vary in their rating severity, and rater incentives (e.g., “get my student a job” vs. “maintain my reputation as a reliable information source”), which might affect ratings, are not transparent. Resumes are not standardized and reveal many skill-irrelevant features of candidates, such as gender, race, and age, which may bias evaluations of candidates (Kessler et al., 2019), and they reflect differential opportunities. They, and their standardized counterpart, biodata, are also susceptible to faking (Law et al., 2002).

Standardized tests are less susceptible to the biases and compromises associated with these alternative measures. As Leonhardt (2023) pointed out, “…perhaps the strongest argument in favor of the tests is that other parts of the admissions process have even larger racial and economic biases.” Chetty et al. (2023) showed that test scores were stronger predictors of outcomes, like attending an elite graduate school or working at a prestigious firm, than were high school grades. They also showed that most of the admissions advantage into prestigious schools of the top 1% income students resulted from higher non-academic ratings (along with legacy preferences and athletic recruitment), not test scores.

But this will be true only as long as the integrity of test scores is safeguarded by a secure process subject to quality control. If score information obtained through standardized testing cannot be relied upon due to security lapses or quality control failures, then the value of standardized testing decreases substantially.

What is the nature of the potential security threats to testing? There are basically three—imposters, informants, and compromised answer keys and prompts—but they appear in many variants. *Imposters* traditionally were persons who took the test for the candidate at the testing site, which has the potential to be made easier with remote testing. *Informants* are like imposters and may be found lurking in the room during home testing, flashing answers to the candidate, out of sight of the computer camera. Another kind of informant is ChatGPT, which has demonstrated tremendous test-taking abilities (Panthier & Gatinel, 2023). Tomorrow’s informants may be communicating to the candidate or taking the test for the candidate via sophisticated communications technologies, exploiting security weaknesses. *Compromised answer keys* were traditionally created from professional or other test-takers’ collective memories of items from the test. Future compromised answer keys may be created with AI tools, such as ChatGPT. Cheating and detecting may remain a cat-and-mouse game as long as the individual incentives to show evidence for skills collide with the greater system's need to ensure the integrity of that evidence.

Approaches to detecting cheating and conducting quality control

Lee et al., (2014) reviewed a variety of research and operational statistical cheating detection methods and quality control tools designed to detect impersonation (imposters), copying (unwitting informants), preknowledge (compromised answer keys), and group collusion (informants). Sinharay (2023) updated Lee et al., (2014) with additional methods. Detection methods include large score difference methods, methods to detect inconsistent performance across test sections that measure related constructs, and examining response times to get clues about whether individuals are responding inconsistently. A widely used method for detecting cheating relies on looking for unusual patterns of responses by groups of examinees, or "unusual agreement between the incorrect answers of two examinees on a multiple-choice test" (Holland, 1996, p. 2), for example, by examinees sitting near each other at a testing site. This is done with the k-index, the Bonferroni-adjusted probability of matching incorrect responses (PMIR; Lewis & Thayer, 1998). With modern communication techniques, examinees may not be near each other, but large groups may still share a circulated answer key and provide a set of responses that match. Statistical methods can identify whether patterns that exactly or nearly match are unusual (Haberman & Lee, 2017); including a patented system for doing so (Haberman et al., 2022).

Long-term monitoring of quality is another approach to evaluating the integrity of the testing process (Lee et al., 2014). Cumulative sum, or CUSUM, charts are commonly used in quality control and they can be applied to testing (Lee & Lewis, 2021). For example, they may help identify an item that, after repeated exposures, no longer elicits the kind of responses it used to, perhaps indicating overexposure. More generally there are a variety of newer statistical methods that can detect abrupt changes in individual items over time. These include harmonic regression (Lee & Haberman, 2013, 2021), time series methods (Lee & von Davier, 2013), and sequential change detection, which can be applied to many applications involving multiple data streams, including testing (Chen et al., 2022). These methods can be applied to identify problematic items that can be excluded from scoring and from the item pool, at least with frequently administered tests. Further developments with these newer methods may involve extending their applicability to some of the newer test forms suggested under Section 3 and may involve adaptation to larger item pools with fewer test-takers per item.

New approaches using AI to detect LLM cheating

The use of ChatGPT and other LLMs on various kinds of high-stakes tests presents new challenges in detecting cheating. Cheating from voice cloning and deep fakes raise new concerns. Hao et al., (in press) suggested a variety of approaches. These include prevention measures, such as the use of additional cameras and test redesign to include items less susceptible to LLM assistance, such as critical thinking and performance-based tasks. Approaches also include detector measures designed to detect ChatGPT contributions to the response, particularly essay responses. Such detectors can be made quite accurate at detecting ChatGPT although false positives are a concern. Hao et al., pointed out that for detectors to be successful all metrics (false positive and true negative rates, equal error rate and contrast samples) have to be considered. *Detectors should be robust against human tweaks to AI-generated text, subgroup bias should be considered, shorter responses are harder to differentiate, and in the end, detectors only can provide probabilistic evidence*. The situation is rapidly changing and open-source LLMs will present challenges (Chakraborty et al., 2023; Liu, Zhang, et al., 2023; Tang et al., 2023).

Scoring—AI scoring methods

Scoring and scoring applications for traditional multiple-choice tests and their variants are well-established scientifically and operationally. Van der Linden's (2018) edited volume presents a comprehensive treatment of the variety of item response theory approaches that can be used for modeling, analysis, scoring, item calibrating, person and model fitting, for tests in various sectors—health, marketing, clinical psychology, and international assessment. There are many other treatments (Ostini & Nering, 2006; Wainer & Thissen 2001) covering approaches applicable to all kinds of tests. These methods are not yet universally applied in operations, nor in research—in many cases, a score is simply the number of correct answers by the test-taker—but increasingly so-called model-based (item response theory) methods are replacing traditional summation (classical) methods in applications in such diverse sectors as organizational behavior (Lang & Tay, 2021), policy, and health (Nguyen et al., 2014).

However, there are several scoring topics that are likely to emerge as an important part of the future of assessments. These are scoring automatically generated items, scoring essays using AI methods, and scoring new novel item types and test-free assessments.

For automatic item generation (AIG) and item difficulty modeling

There are several approaches for automatic item generation (AIG; see chapters in Gierl & Haladyna, 2013, particularly Sinharay & Johnson, 2013; Irvine & Kyllonen, 2002). The radicals and incidentals approach involves building items from a set of factors or dimensions by varying the values on those dimensions. Factors that influence difficulty are called radicals, those that do not are called incidentals. The factors are based on a cognitive analysis of the domain. This is the approach taken by Embretson (1999) and Kyllonen et al., (2019) and is ideally suited to algorithmic items such as progressive matrices or number series fluid reasoning items. It uses the linear logistic test model (LLTM) and extensions to model the data and as the basis for scoring.

The other approach is the item-model approach, which may be called slot-filler, in which parts of the model item (e.g., some quantities in an arithmetic word problem) are treated as slots that have a set of associated potential fillers. This approach was taken by Bejar (2002; Graf & Fife, 2012) and is ideally suited for mathematics or physics word problems. Johnson and Sinharay (2005) reviewed the approaches proposed for scoring these and suggested that a simple model called identical siblings (ISM) does a reasonably good job of estimating test-takers' abilities by assuming that all items made from the same item model are the same, regardless of the fillers. However, relaxing this assumption, which is done in the related-siblings model (RSM; Glas & van der Linden, 2003) and linear item cloning model (LICM; Geerlings et al., 2011) potentially applies to a much broader class of AIG assessments by allowing the inclusion of collateral information and enabling a more rigorous statistical analysis.

Scoring essays and other hard-to-score tasks

Automated, machine-scoring of essays is now well-established in operational scoring. Current versions are based on statistical learning methods, primarily multiple regression and other predictive AI approaches, such as random forests and gradient-boosting machines (Madnani & Cahill, 2018; Rupp et al., 2018; Shermis & Burstein, 2013). Automated essay scoring is nearly as accurate[7] as human scoring and it brings advantages in the avoidance of human biases related to rater fatigue, severity and leniency, drift, time-of-day, and halo effects (Williamson et al., 2012). On the other hand, there is a perception that automated scoring is a black box, which might have its own biases, engendering a lack of trust from test-takers (Kumar & Boulanger, 2020).

Deep learning models and LLMs are beginning to be used in essay scoring and in the evaluation of other hard-to-score tasks. They have the potential to increase accuracy and to provide better explanations to test-takers about the strengths and weaknesses of their assessment work products (Kumar & Boulanger, 2020). Hao et al., (in press) discussed several applications of LLMs to automated scoring. One investigation found extremely high correlations between human ratings and AI-based automated scoring of student responses from eight countries, six different languages, across six items from TIMSS 2019 (Jung et al., 2023). The relationship was particularly strong when the system was trained on ChatGPT-translated responses. The other was an application of convolution neural networks to scoring TIMSS 2019 graphical response items that can appear on science mathematics tests finding high levels of accuracy and identification of some human rating biases (von Davier et al., 2022). This work is promising but in its early stages and there is likely to be a flurry of activity applying LLMs to scoring of short answer, essay, graphical-response, and other hard-to-score tasks in the coming years. A major challenge in this work will be the avoidance of bias in AI-powered scoring models, a topic addressed in Duolingo English Test's Responsible AI Standards (Burstein, 2024; Johnson, 2024). Johnson et al. (2022) discussed the example of "rubric-irrelevant" response features associated with performance, such as writing style, length of response,

[7] In the automated essay scoring literature, accuracy, or exact match, is commonly measured as exact agreement between two raters or a machine and human score. Accuracy can also be normalized to a baseline of random chance as in the kappa, (linear) weighted kappa (Cohen, 1968), and quadratic weighted kappa (which penalizes discrepancies beyond a linear penalty) measures. A common approach is to consider degradation from human human score agreement. Using these measures, Williamson et al., (2012) reported that for many types of essays there was minimal degradation from human agreement, and in fact, "it is relatively common to observe automated–human agreements that are higher than the human–human agreements" (p. 8). More recent studies using transformer-based approaches (Ormerod et al., 2021) report above-human-level performance.

and typos, that happen also to be associated with a demographic variable. Solutions for some of these kinds of issues are beginning to be proposed (Johnson & McCaffrey, 2023), and AI bias in scoring is likely to remain a promising area of investigation. At this point, LLM-assisted item development and response scoring remain research topics requiring active human-in-the-loop participation as LLM hallucinations and AI bias preclude fully autonomous systems.

Scoring testless assessments

Testless assessments might be defined as assessments of skills based on behaviors or behavioral traces that are not connected to an explicit test. This includes conversations during problem-solving or learning, employment interviews (Emerson et al., 2022), the actions taken when freely exploring a game or microworld environment, or even resume items; L data in Cattell's (1965) terminology. This is a disparate set of activities, and consequently a wide variety of approaches have been taken to model behavior in these environments. For the most part, these approaches have not been connected to the psychometrics literature. Methods have ranged from studying on- versus off-task behavior patterns in school children by gender and across time (Godwin et al., 2016) to exploring keystroke patterns on standardized tests with exploratory items (He et al., 2019), examining questionnaire item skipping as an indication of disengagement (Hitt et al., 2016; Mignogna et al., 2023), to characterizing conversations by coding them using machine learning methods (Kyllonen et al., 2023). There is likely to be significant further development in this area that will use LLM approaches for classification and other kinds of data exploration.

Fairness

Fairness, or the minimization of bias, is considered "an overriding, foundational concern" in testing (AERA et al., 2014, p. 49) because it affects the justification or validity of a test score interpretation. An interpreter of a test score should be able to assume that the test measures the same underlying construct regardless of characteristics of the test-taker, such as disability or language status, or culture or language background.

> A test that is fair within the meaning of the *Standards* reflects the same construct(s) for all test-takers, and scores from it have the same meaning for all individuals in the intended population; a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct. ...characteristics of all individuals in the intended population, including those associated with race, ethnicity, gender, age socioeconomic status, or linguistic or cultural background, must be considered throughout all stages of development, administration, scoring, interpretation, and use so that barriers to fair assessment can be reduced. (AERA, 2014, p. 50).

This notion of test fairness can be addressed at the item writing stage through the consideration of guidelines "designed to eliminate symbols, language, and content that are generally regarded as sexist, racist, or offensive, except when necessary to meet the purpose of the product or service" (ETS, 2014, p. 21; see also ETS, 2022), and through the review process that checks for accessibility and fairness related to item content. Fairness also is addressed through statistical analysis of item responses that examine the degree to which the test measures the same construct for different groups based on gender, race, language, culture, and other factors. Statistical methods can be used to identify an item that does not behave the same way in two groups, due, for example, to differential familiarity with a word across gender groups (e.g., a sports term) or culture groups (e.g., a food item). Treatments of this topic are found in Millsap (2011).

A second definition of fairness is a concern in employment testing and based on the selection rate for different groups of test-takers defined by gender, race, and age. If the selection procedure has an "adverse impact," meaning that it screens out members of a "protected group" at a rate higher than the most favored group, then the employer could violate the "Uniform Guidelines on Employee Selection Procedures" (a similar concept in the EU is "indirect discrimination"), and subject to legal enforcement actions by the Equal Employment Opportunity Commission.

Bennett (2023), in line with Solana-Flores (2019) and Sireci (2020), argued that beyond these definitions, the basic premises of educational assessment must be rethought because of opposition to traditional standardized tests related to a "perception that tests represent a worldview no longer suited to the pluralistic society we are rapidly becoming" (pp. 17-18). His proposal is to design "socioculturally responsive" assessments (CRAs) by changing content to be culturally relevant, providing population-specific assessment, adapting the assessment to student characteristics, and encouraging learner agency (O'Dwyer et al., 2023). Bennett hypothesized that culturally relevant problems on a test would increase test-takers' identification with the assessment, engagement and motivation, activation of prior knowledge and consequently their test performance and confidence and sense of efficacy. From the standpoint of cross-cultural research (see Contextualization and Personalization), Bennett is proposing an adaptation step, perhaps more extensive than typical adaptations, to assessments. Walker et al. (2023) proposed provisional principles for designing CRAs that consider students' background characteristics like "beliefs, values, and ethics; their lived experiences; and everything that affects how they learn and behave and communicate" (p. 1) Dobrescu et al. and Kukea Shultz and Englert (2021) have field tested CRAs, but neither have formally shown that the CRA is equivalent to the non-CRA version of the test.

Sinharay and Johnson (in press) addressed this limit and proposed a statistical and psychometrics framework for analyzing data from CRAs, which obtains "equivalent evidence about examinees from alternative, not-surface-equivalent, forms of tasks" (Mislevy et al., 2018; see Feuer et al., 1999, for discussion of equivalence). Sinharay and Johnson (in press) achieve this by pairing items across two forms, one designed for the reference group (RGV), and another adapted, as per Bennett (2023), for the focal group (FGV). The researcher then establishes forms equivalence through expert judgments, and within-form psychometric analyses (difficulty, discrimination, reliability, factor structure, differential item functioning (DIF), and test characteristic curves), then, having done that, produces form-specific scores that can be treated as interchangeable by policy. However, examinees can also receive scores on both forms (RGV, FGV), which therefore measure "within-context" and "out-of-context" abilities. In a simulation study testing varying designs relating to which group receives which form, Sinharay and Johnson (in press) found that as long as some items were essentially common across forms, it may be possible to treat scores from the two forms as comparable.

Other approaches besides creating entirely different forms may be possible for addressing the general issue of bias on tests related to the cultural background of test-takers. For example, De Boeck and Cho (2021) presented an alternative category of DIF based on the statistical concept of treating person and item effects as random rather than fixed effects and using explanatory covariates "to explain the variation while allowing DIF to pervade subsets of items and even the whole test, if helpful to understand the item responses." De Boeck (2023) used the example of participants varying in their familiarity with stimulus material, where that variation was related to performance (i.e., an explanatory covariate). Such a covariate, such as cultural familiarity or opportunity to learn, if operationalized, could likewise serve as an explanatory covariate that could help explain item responses and serve as the basis for test scoring that accounted for cultural familiarity or opportunity to learn.

Conclusions for Section 4: Operations breakthroughs

Testing operations, which includes considerations of the purpose of the test and the administrative conditions and constraints, along with the item development, test assembly, security, quality control, scoring, and test evaluation, are the heart of the testing industry. There are many challenging issues in operations associated with making tests valid, reliable, fair, and useful to the test-taker and other stakeholders. It is likely that advances in technology, particularly LLMs and other AI technology will have a dramatic effect on testing operations, as technology has since the beginning of testing. We are likely to see significant advances in efficiency and quality related to how tests are developed, assembled and scored, made secure and made fair so that all test-takers can see value in tests and can be confident that inferences drawn based on test scores are appropriate and justified.

SECTION 05.

Feedback:

learning science-driven insights and action plans for test-takers

In this section, we review various efforts to examine how assessment can facilitate learning. These include efforts to combine assessment and learning, formative assessment, the testing effect, tutoring, and intelligent tutoring systems. We also review diagnostic assessments, process analyses, what we know about what works, and the effects of feedback, and conclude with a review of principles of learning. All of these have implications for how we can best provide useful information to test-takers to facilitate the achievement of their educational and career goals.

Setting the stage

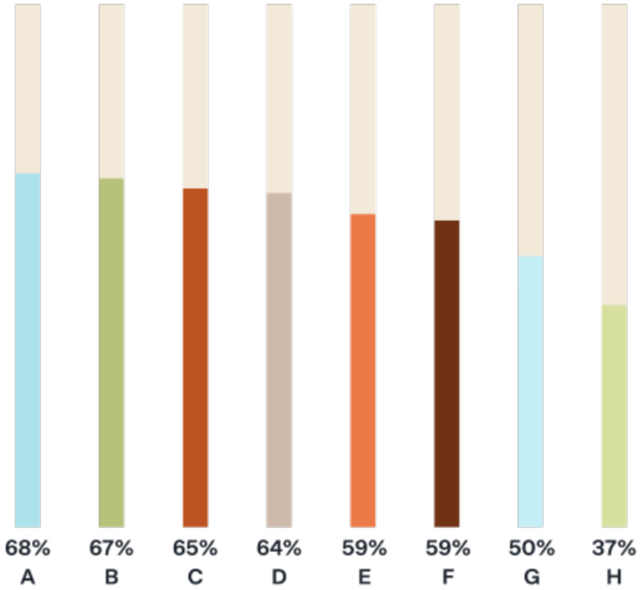
Test users are increasingly looking for additional information from examinations beyond whether they passed, were accepted, or received an award. The ETS Human Progress Study (2023a) found that, if given an assessment with career guidance, respondents would more likely be motivated to acquire new skills, be prepared to face challenges, feel recognized for their personal performance, feel confident in their abilities and in pursuing new job opportunities, and other positive sentiments (Figure 4).

> “In the future, we will be able to develop a personalized assessment, based on individual abilities and aspirations, which would be a great step forward.”
>
> **Joana Lenkova,**
> **Futurist and Strategist,**
> **Futures Forward**

FIGURE 4.

PERCENTAGE OF RESPONDENTS REPORTING INCREASED LIKELIHOOD OF VARIOUS SENTIMENTS RESULTING FROM SKILLS ASSESSMENTS.

A	Be motivated to acquire new skills or knowledge
B	Feel more confident in my abilities
C	Feel more confident pursuing new employment or job opportunities
D	Feel prepared to face challenges
E	Feel recognized for my personal performance
F	See a direct link between skill development and career advancement
G	Feel less stressed about the future
H	Stay with my current employer



SOURCE: ETS Human Progress Study (2023a). Questions: "If you were able to take skills assessments and receive guidance as a pathway for career growth, would you be more or less likely to do or feel any of the following?" (Less likely/No change/More likely).

In many contexts, test scores are returned with norms, benchmarks, and descriptive information that helps test-takers interpret their scores, but more could be done to provide useful feedback that might help test-takers advance toward their education and career goals. The purpose of this section is to address this issue from the standpoint of what might be done in the future to provide useful, actionable information to test-takers based on principles of learning science, and findings from the formative testing, assessment-for-learning, testing effect, feedback, and tutoring literatures.

Diagnostic testing has long promised to provide deeper insights into test-taker knowledge and skills and to provide the foundation for better, personalized feedback to test-takers. Yet, despite significant advances in the psychometric modeling underlying diagnostic testing (Rupp et al., 2010), the reality has not always lived up to the promise.

One method that might contribute to positive change in diagnosis and in our ability to provide useful information back to the test-taker is process analysis. Process analysis involves looking beyond item responses to interpreting additional information such as response times, specific test-taker actions, and in the case of collaborative problem-solving or collaborative learning, the conversations occurring during problem-solving. Process analysis potentially provides raw data for providing useful insights into what learners know and can do or do not know and cannot do. Process data, perhaps coupled with diagnostic modeling, might provide a greater foundation for providing prescriptive interpretations of test-taker knowledge and understanding of a topic.

Intelligent tutoring or adaptive training is one area where process analysis is being done (Greiff et al., 2017). The log file containing learner actions (process data) is analyzed in real-time to form a dynamic model of the learner's knowledge, which guides the selection of instruction and the process of estimating learner proficiency. Process and response data contribute to the assessment of student knowledge, and feedback is provided throughout. Human tutors similarly test students' knowledge and understanding through questioning and adapting their instruction accordingly. Analyzing how feedback is used in the tutoring context, human (Nickow et al., 2020; VanLehn, 2011) or computer (Duolingo Team, 2023; Kahn, 2023; Sottilare et al., 2018), might inform how feedback can be used in assessment generally.

Finally, principles of learning ought to serve as the foundation for efforts to provide useful feedback to test-takers. Feedback is a form of instruction, and it is useful to review what we know about what works in learning, which is commonly encapsulated in principles of learning.

Paradigms for combining assessment and learning

The organization of this section is as follows. First, we review paradigms that combine assessment and learning. These include formative assessment; its identical twin, assessment for learning; the testing effect from memory research in cognitive psychology; and tutoring, human and machine. Next, we review studies of the effects of feedback in instruction. We next review diagnostic assessment and process analysis. We follow this with a review of the principles of learning. We conclude with a discussion of how feedback given during testing can contribute to positive outcomes for individuals, can address equity, and can benefit society more broadly.

Formative assessment (assessment for learning)

There are many concepts and associated literature relating to how assessment can improve learning. One of these is *formative assessments*, which is a broad concept with many, diverse definitions (Bennett, 2011). Xuan et al., (2022, their Appendix A) identified 19 of them from various studies; and organized them by what, why, when, who, and how questions. To informally characterize Xuan et al.'s (2022) list of definitions, *formative assessment is an assessment (or a process or a tool) to improve competence (or adapt instruction or identify where learners are and where they are going) during instruction (or while teaching) involving teachers (or students or peers) using some approach (approaches are too variable to succinctly characterize)*. Shepard et al.'s (2017, p. 275) definition is the simplest, that formative assessments are ones to improve teaching and learning during the instructional process,) and Black and Wiliam's (1998, p. 7-8) definition is perhaps the most influential that formative assessments are assessment activities that provide information to be used as feedback to adapt teaching and learning to meet students' needs). Related to formative assessment is assessment for learning (AfL), assessment as learning, formative evaluations, and curriculum-based assessment. But in their meta-analyses, Xuan et al., (2022) and Klute et al., (2017) treated these all as indistinguishable (Xuan et al., also included "feedback" in their meta-analysis of formative assessment). There are no comparable lists of "AfL" or "assessment as learning" definitions, so it is convenient to treat the two concepts as synonymous.

There have been several meta-analyses on the effects of formative assessment, which, not surprisingly, given the variability in definitions, have yielded variable estimates of their effects. After Fuch and Fuch's (1986) initial findings of an effect size of.7[8], Kingston and Nash (2011) found more moderate effect sizes of .32, .17, and .09 for English language arts, mathematics, and science, respectively, the difference being attributed to differences in study inclusion based on rigorousness. Klute et al., (2017) found effect sizes of .36 for math, .22 for reading, .21 for writing, and a difference based on whether formative assessment was student-directed, with an effect size of .45, in math, or other-directed with a .30, in math. Student-directed meant that students worked in groups on material with a protocol but without a teacher; other-directed meant that a teacher supervised and adapted lessons. In reading, other-directed led to greater gains than student-directed assessment. A limitation of the studies comparing student- versus other-directed assessment was that the number of them was small and the interventions were different and so the observed differences might have been due to intervention features other than the student- vs. other-directed aspect. Xuan et al.'s (2022) meta-analysis added the findings that teacher-student collaborative formative assessments were more effective than teacher-initiated ones, and that differentiated instruction, or adapting instruction based on results from the assessment, was more effective than non-differentiated instruction, and that there were Anglophone-Confucian-heritage cultural differences, with higher effect sizes for the latter.

Testing effect

The *testing effect* is a term to describe the benefits to the memory of being tested—specifically, being tested on a concept can improve learning of that concept. The testing effect emerged from the human memory literature in cognitive psychology (Karpicke & Blunt, 2011). The basic idea is that learning can be divided into initial instruction (exposure), followed by studying (or practice); and then final testing. The empirical finding is that if some interim testing is substituted for some of the studying, then final testing will show greater memory for the material than without the substitution.

[8] In this article we refer to effect sizes, which are indications of the strength of the manipulation on (or relationship with) outcomes. Classic rules of thumb proposed by Cohen (1992) are that small, medium, and large effect sizes correspond to effect size values greater than .20, .50, and .80, respectively.

This is true even if the comparison condition, the studying phase, involves active learning such as *memory elaboration*, which is known to produce greater memory gains than simple rehearsal. One way to think about why testing per se produces gains relative to studying, is that interim testing provides an opportunity for retrieval practice, particularly, but not only if the interim (and final) testing is recall testing. Retrieval practice is valuable during later testing because later testing involves retrieval. Hence another term for the testing effect phenomenon is *practice tests* or *practice testing* to convey the idea that what a learner is doing when testing is practicing test taking (Adesope et al., 2017).

There have been several meta-analyses supporting findings on the testing effect since Bangert-Drowns et al.'s (1991) initial study. Rowland (2014), focusing on laboratory studies, examined the evidence for a broad variety of theoretical explanations of the testing effect. He found a .50 effect size for the testing effect when compared to restudying. He also found that the testing effect was larger for recall, but still present for recognition testing, that it operated over both short and long intervals, and that it operated for both verbal and nonverbal materials. The testing effect is not confined to laboratory studies. Phelps' (2012) meta-analysis defined the testing effect much more broadly to include a large set of studies conducted over the past century on the effect of testing generally, finding effect sizes ranging from .55 to .88. Adesope et al., (2017) limited the scope, compared to Phelps, including only quantitative, low-stakes studies, examining 272 effects from 118 experiments, finding an average effect size of .61 (.51 when compared to a control condition of studying per se; .93 when the control condition was unrelated to the test). They also found that multiple-choice testing as a treatment (.70) gave a bigger boost than short-answer testing (.48) with both together even higher (.80); that single rather than more than one practice test was best and that the effect occurred similarly in lab and classroom settings, and across primary, secondary, and postsecondary settings.

Besides the testing effect, Roediger et al., (2011) identified empirically based benefits of testing and their applications to education. Direct benefits are that retrieval practice increases retention of the material (the testing effect), and to related material, and it facilitates transfer to new situations. Open-ended assessments also help students organize information. There are also indirect benefits: frequent testing motivates students to study more, discover gaps in one's knowledge (due to explicit or implicit feedback, specifically, knowledge of results from testing), and focus efforts on more difficult material. Roediger et al., argued for more self-testing and for frequent quizzing.

Tutoring

Human tutoring, either one-on-one or small group (five students or fewer), is considered to be among the most effective forms of instruction. Bloom (1984) originally presented evidence (from three studies) that tutoring from a good tutor provided a 2 standard deviation improvement (i.e., effect size of 2.0) over conventional instruction (and about half that over mastery learning with formative assessment). He argued that one-on-one tutoring is too costly but that a societal goal should be to determine how to accomplish the benefits of tutoring but with more practical and realistic methods, which he referred to as the "2 sigma" problem.

Dietrichson et al.'s (2017) meta-analysis with 36 studies similarly concluded that tutoring (along with feedback and progress monitoring and cooperative learning, although these had slightly lower effect sizes) was the most powerful academic intervention identified for its effects on standardized achievement test scores (of 14 intervention types) for the low socioeconomic population they studied, although with a more modest but still substantial effect size estimate of .36 or .32 for feedback and .22 for cooperative learning. This is compared to average intervention effect sizes of .09 for reading and .08 for mathematics. Dietrichson et al., (2017) screened for study rigorousness (e.g., treatment-control designs, most of which, 76%, were randomized control trials and use of standardized achievement tests as outcomes (to avoid intervention content contamination bias), perhaps explaining some of the differences between their and Bloom's (1984) estimates of the size of the tutoring effect.

Whereas Dietrichson et al., (2017) focused on interventions, Nicknow et al., (2020) focused directly on tutoring per se, examining 96 studies to determine impacts and the effects of program characteristics and context. They found an effect size estimate of .37, concluding that "tutoring programs rank among the most flexible and potentially transformative learning program types available at the PreK-12 levels." They found the effects to be larger when tutoring was done by teachers and paraprofessionals than by parents, larger in earlier grades, and larger when conducted

[9] VanLehn (2011) argued that Bloom's (1984) exit conditions (mastery levels needed to move to the next lesson varied, and so Bloom (1984) actually presented evidence for the effects of mastery).

in school rather than after school. They suggest that after-school parent tutoring is harder to control for implementation.

Why is tutoring—the most powerful educational intervention yet identified—so effective? Nicknow et al., (2020) proposed several possibilities. Because tutoring is typically used to supplement classroom instruction, *tutoring might simply provide more study time*. Another possibility is that *tutoring customizes learning to the right level for the student, which is also accomplished with tracking and class size reduction to lesser degrees*. Another is that *tutoring promotes engagement and enables rapid feedback, stimulating more student effort*. Still another is the *human connection, the mentorship relationship*.

VanLehn (2011) similarly proposed a potential mechanism by which tutoring might affect learning outcomes by focusing on what humans can do better than computer tutors. Among the plausible hypotheses that he failed to find good research support for were these: 1) human tutors develop a detailed diagnostic model of the student's knowledge and misunderstandings; however, there seems to be little support in the empirical literature that human tutors actually do this; 2) tutors select just the right task for students given what students need; although this is likely true, computer tutors can do the same, and so this would not be a human advantage; 3) human tutors are thought to have the ability to use sophisticated tutoring strategies, but studies show that humans tend not to use sophisticated tutoring strategies; 4) humans have deep knowledge of the subject matter and can bring in related ideas, but studies show this tends not to happen, or if related knowledge is brought in it does not affect outcomes; 5) like Nicknow et al., (2020), VanLehn (2011) proposed increased motivation due to "the warm body effect" or the tutor's ability to give praise, but he found little support for either hypothesis.

VanLehn (2011) did find support for several hypotheses: 1) human tutors provide feedback and hints immediately when needed; 2) human tutors scaffold students' reasoning (provide "guided prompting"); and 3) tutors encourage interactive and constructive (vs. active and passive) behavior which makes for increased learning. VanLehn (2011) also suggested that these latter three hypotheses, which were supported in the empirical literature, were consistent with Chi and Wylie's (2014) Interactive, Constructive, Active, and Passive (ICAP) framework. This framework suggests that engagement behaviors can be categorized into the four ICAP modes, including interactive, constructive, active, and passive and that learning increases as students become more engaged with the learning materials in a passive to active to constructive to interactive continuum. Interactive learning is the pinnacle of engagement and thus learning.

Intelligent tutoring (adaptive instructional) systems

Intelligent tutoring systems (ITSs), or adaptive instructional systems (AIS), are methods for using computers as tutors, an effort to solve Bloom's (1984) 2 sigma problem. There is a large literature on ITSs and an 11-volume series (Sinatra et al., 2023), including strengths-weaknesses-opportunities-threats (SWOT) analyses on all aspects of ITSs (Goldberg & Sinatra, 2023) The traditional architecture of an ITS consists of a learner model, a domain or curriculum model, and a pedagogical model. The learner model represents the learner's current relevant knowledge and skills levels as well as current state. The domain model represents the curriculum or instruction to be taught and includes rules for selecting domain content (adaptive sequencing). The pedagogical model identifies when feedback is needed based on the learner's performance (adaptive feedback), which is integral to ITS architectures. An alternative and perhaps more streamlined definition is that an ITS is a system that provides personalized prompts, hints, and support feedback during rather than only after problem-solving (VanLehn, 2011).

Numerous meta-analyses have been conducted with ITSs. Fletcher and Kulik (2015) estimated an effect size of .66, compared to conventional instruction, but their analysis included non-experimental and lab studies, which tend to show larger effects than field studies. Ma et al., (2014) and Nesbit et al., (2014) also found evidence for ITS efficacy but with more modest effect sizes (.43). Steenbergen-Hu and Cooper (2013) found negligible effects on mathematics learning, perhaps because they included less effective instructional systems that did not meet an ITS definition. The mixed results suggest that it is important to consider the core components of ITSs, determine which are critical, and consider implementation issues.

There seems to be some evidence that adaptivity is an important component. Adaptive instructional systems make use of learner models to implement personalization (e.g., adaptive feedback, adaptive sequencing of tasks or activities). These learner models can include information about the learner's cognitive, metacognitive, affective, personality, social and perceptual attributes (Abyaa et al., 2019; Shute & Zapata-Rivera, 2012). Learner models can also be made available to learners, teachers, and other audiences to support metacognitive processes, collaboration, navigation, trust, and accuracy of the model (Bull & Kay, 2016). The type of information and mechanisms used to share learner model information depends on the needs, knowledge, and attitudes of each audience (Zapata-Rivera, 2020).

Grain size is defined as "the amount of reasoning required of participants between opportunities to interact" (VanLehn, 2011, p. 202). VanLehn proposed a continuum of granularity, from coarse, answer-based tutoring, in which feedback comes only after an answer is provided (as in adaptive testing), through step-based tutoring, in which feedback comes after a problem-solving step (such as asking for a hint), through substep-based tutoring, in which feedback and scaffolding are provided at a finer level than the steps taken during problem-solving, to human tutoring, where the tutor can interrupt at any time. VanLehn proposed an interaction granularity hypothesis, that tutoring is effective to the degree to which it provides feedback within or after a problem-solving step vs. after an answer is entered. VanLehn found evidence that step-based tutoring was as effective as substep, and therefore was an optimal grain size.

Feedback is also critical. Adaptive feedback can vary by the amount of information provided (e.g., verification feedback, hints, elaborated feedback), the timing of the feedback (e.g., immediate, delayed), and the goal of the feedback (e.g., informing immediate next steps, providing guidance on progress made toward achieving an instructional goal; Shute, 2008). Adaptive features (e.g., personalized feedback) can be provided at the macro level, where the best task to achieve an instructional goal is selected and the micro level, where different aspects of the current task including the level of feedback are adjusted (VanLehn et al., 2007).

Diagnostic assessment and process analysis

Cognitive diagnostic modeling (CDM) is a set of methods for modeling responses to test items or tasks when those items are coded by features indicating the cognitive processing requirements associated with item solving. The motivation for modeling response data using CDMs is to reveal learners' underlying information processing, thereby gaining an understanding of the learner based on the pattern of items answered correctly and incorrectly, and using the cognitive requirements of the items to infer what they know and do not know and where there might be misconceptions based on that pattern. The promise of cognitive diagnostic modeling has been to reveal characteristics of learners' problem-solving for diagnostic purposes, so that instruction and feedback can be personalized to the learner. The motivation and approach behind CDM is similar to the motivation and approach behind student modeling in the ITS literature, which also is concerned with personalization, but until recently their histories were independent: CDM is a branch of psychometrics (von Davier, 2010) and ITS student modeling grew entirely out of the learning literature in cognitive psychology (Corbett & Anderson, 1994) adopting a method called *knowledge tracing* for student modeling (Liu, Kell, et al., 2023).

There is a long history and large literature on cognitive diagnostic modeling and assessment (Rupp et al., 2010). Recent efforts have incorporated process data, such as response times (Zhan et al., 2018) as a way to gain greater insight into individuals' learning processes. There also have been attempts to link the CDM and ITS student modeling literature (Wang et al., 2018). One approach is to use CDMs for Bayesian knowledge tracing (BKT), a student modeling method used in the ITS literature based on hidden Markov models (HMM; Wang et al., 2018, 2020). Wang et al., (2018) combined a BKT HMM with a CDM framework, enabling the tracking of the growth of multiple skills and accommodating covariates to model the HMM skill transitions. There is a rapidly growing area of research and there are likely to be continued developments promising increasingly accurate, interpretable, and actionable cognitive diagnoses to enhance personalization (Pu et al., 2021; Wang et al., 2020).

Feedback

The discussion thus far implicates feedback as a critical core component of many education interventions—formative assessment, the testing effect and human and machine tutoring. There is also an independent literature on the effects of feedback per se on educational outcomes. In an early study demonstrating the potential of computer-based instruction, Azevedo et al., (1995) found that feedback in computer-based instruction produced an effect size of .80 on immediate achievement posttests and .35 on delayed posttests. Hattie and Timperley (2007) estimated an effect size of .79, and more recently, Wisniewski et al., (2020), using stricter exclusion rules, estimated an effect size of .48, but with significant heterogeneity. Feedback effects were larger for cognitive and motor skills compared to motivational and behavioral skills. Feedback also was more effective the more information it contained—it was most beneficial when it helped students understand what mistakes they made, why they made them, and how to avoid them in the future. Timing of the feedback was also found to be important (Hattie & Timperley, 2007). Immediate feedback is often more effective, but delayed feedback may be more effective when learners are engaged in complex tasks (e.g., Attali & van der Kleij, 2017; Fyfe et al., 2021; Hattie, 2009).

Effective feedback should consider instructional context, nature of the task, and characteristics of the learner (Shute, 2008). What is most effective can depend on the situation. Panadero and Lipnevich (2022) provided an integrative typology of feedback likely to be effective in different situations. The typology categorizes feedback by *content* (e.g., verification, elaborative), *function* (e.g., support learning, instill motivation, foster mastery-orientation), *presentation* (e.g., immediacy, frequency, adaptivity to the learner's progress, number of modalities used to convey the feedback), and *source* (e.g., teacher, peer, self, computer).

Informative feedback affects both achievement and motivational variables such as engagement, effort, persistence, and satisfaction (Narciss, 2004). Shute (2008) argued that effective instructional feedback should have several characteristics. It should appear unbiased (Kluger & DeNisi, 1996; Panadero, 2023) and focus on the task and not the learner (Fyfe et al., 2023). It should be elaborated to engage the learner and lead to long-term learning in the presence of misconceptions (Attali & van der Kleij, 2017) yet presented in a manageable, specific, and clear format (Moreno, 2004). It should be administered only after the learner has attempted a learning task (Hattie & Gan, 2011). It should be provided to promote continuous learning while reducing some degree of mismatch between the learner's current performance and the intended learning outcomes (Leenknecht et al., 2019).

Implications for the design of innovative assessments

Effective feedback is a tool to direct learners on how to improve their learning and use resources that can help them make improvements (Hattie & Timperley, 2007). Understanding characteristics of effective feedback can inform the design of digital learning and assessment systems that offer the benefits of human-to-human feedback but at a much greater scale. Such knowledge can be used to structure feedback features within innovative assessment systems, ensuring that all learners receive a diverse range of feedback types that align with specific learning goals.

Future digital assessment design should consider methods for providing *personalized feedback that caters to each learner's strengths and weaknesses* (Lipnevich & Panadero, 2022) to lead to more effective learning experiences. To ensure all students can access and have opportunities to engage in more equitable learning experiences, digital assessments must provide feedback that is clear, accessible, and accommodating of diverse student needs. Well-designed personalized *feedback can also be highly motivational*, providing students not just information to improve learning, but also propagating greater interest and value associated with a learning task (Narciss et al., 2014). Depending on the design of assessment, feedback can also engage learners through dialogical interactions with teachers, peers, or simulated agents, making the learning experience more interactive, collaborative, and engaging.

Understanding and including feedback for learners in digital assessments should lead to innovations and improvements in the future of assessments in several ways. Feedback-seeking behaviors within digital learning platforms can themselves serve as indicators of behaviors associated with learning self-regulated learning. *For example, analysis of clickstream data related to these behaviors provides insights into how students are managing their learning processes, seeking guidance, and adapting their strategies based on feedback provided to them within a digital learning and assessment platform* (e.g., Aguilar et al., 2021; Bernacki, 2017; Lu et al., 2017; Ober et al., 2023; Tenison & Sparks, 2023). This understanding can inform the design of assessments that promote self-regulated learning and guide students toward more effective study habits. Support for learners could be customized based on individual students' receptiveness to feedback, their feedback-driven behaviors, and their specific areas for improvement. Leveraging multimodal data sources can provide a more holistic view of students' learning behaviors, leading to more effective and personalized interventions (Lehman et al., 2018; Sparks et al., 2023); Zapata-Rivera et al., 2020). These advancements have the potential to enhance the overall learning experience and contribute to better educational outcomes in digital learning environments.

Principles of learning

The foundation for providing feedback to test-takers should rest on a foundation of learning principles. The preceding section provides such principles as they have emerged to support the generation and delivery of feedback. But it is also useful to consider broader principles of learning that have emerged over the past century. This is a vast literature but there have been several useful syntheses that may be especially appropriate for the context of the future of assessments. Thorndike proposed three laws of learning—the laws of effect (reinforcement), exercise (practice), and readiness, which have remained viable. The American Psychological Association's (2018) top 20 principles from psychology for preK-12 teaching and learning includes principles related to thinking and learning, motivation, social and emotional context, classroom management, and assessing student progress. Carnegie Mellon University's Eberly Center (2024) proposed a set of seven principles underlying effective learning: prior knowledge can help or hinder learning, knowledge organization influences learning, motivation governs learning behaviors, combining and practicing skill components and goal-directed practice with feedback is important, social and emotional as well as intellectual aspects are important, and self-monitoring and adjustment is important to becoming a self-directed learner. Schwartz et al., (2016) provided an evidence-based summary of 26 learning principles, designed to be used by educators.

Bjork and Bjork (2011) put the spotlight on a general set of principles falling under the heading of *desirable difficulties*—conditions of learning that create difficulty but lead to more durable and flexible learning. These include varying the conditions of practice, spacing study and practice sessions rather than cramming them together (before the test), interleaving (versus blocking) instruction on tasks that are to be learned as part of a larger whole, and the generation and closely related testing effects (reviewed in a previous section). For each of these, the easy condition might lead to short-term gains, but the difficult condition leads to longer-term gains and the ability to use the acquired knowledge more flexibly, hence, desirable difficulty.

The National Research Council (2000) and National Academies of Science, Engineering and Medicine (2018) have produced a comprehensive two-volume series *(How People Learn, How People Learn II)* summarizing learning principles applicable across the lifespan and school and work settings from diverse disciplines. A number of conclusions are drawn from topics covering culture, types of learning, knowledge and reasoning, motivation, school learning, technology, and learning across the lifespan. Recommendations are provided for future research on the importance of learning contexts and technology in learning. Principles of learning captured in these two volumes and elsewhere can be used to guide the development of feedback that can be administered in the context of assessment.

Conclusions for Section 5: Feedback

The problem addressed in this section is that testing often asks a lot from test-takers, in time, effort, and expense, and often does not reciprocate in providing much direct educational value in return. A question is what kind of value can a test provide to the test-taker? In this section, we reviewed several ways in which assessment and testing is used to provide useful information or aid in skill acquisition to test-takers or assessment targets. We also provided evidence-based estimates of the value of the information, or the aid testing can provide. We focused on information that goes beyond the test score and the norms and benchmarks that are often provided, although normative and interpretive information provides significant value.

Formative assessment, which involves using testing as an integral part of the instructional process, although implemented in wildly varying ways, was shown to provide a significant, positive effect on learning. The testing effect, or testing practice, which simply substitutes some of the learner's studying time with time instead spent testing, was similarly shown to have strong positive effects on learning outcomes. Human tutoring has been found to be among the most powerful educational interventions. Computer-based intelligent tutoring, or adaptive instruction, similarly is a powerful intervention. The reason tutoring, human or machine, is so powerful is not completely understood, but there is some evidence that providing feedback, guided prompting, and encouraging interaction and constructive behavior are important components. Tutoring also performs cognitive diagnosis of the learner and testing can similarly do so. Increasingly sophisticated cognitive diagnosis modeling that takes advantage of AI advances and incorporates more process behavior from the test-taker into the learner model promises to provide useful assistance to learners in personalizing instruction. Feedback, too, was found to be a powerful means to improve learning through personalization. Much is known about what kinds of feedback are most effective, and the use of generative AI to provide useful feedback to learners and students is a promising new direction. Finally, we know much more about the learning process itself and what works to produce enhanced outcomes than we did even a couple of decades ago. Adhering to evidence-based learning principles in the formulation of feedback, instruction, and guidance to learners will enhance the value of assessments significantly.

Thus, assessment can be a two-way street in which learners provide information to teachers or policymakers on their skill levels, and the learner also receives something valuable in return—guidance on where to go next in the learning journey to close gaps between the learning objective and the current skill level, increased skill, and a sense of autonomy, competence, and belonging. The provision of feedback, appropriately designed and personalized, can serve the goals of equity in education and promote learning and performance for all learners.

SECTION 06.

Summary and conclusions

The purpose of this paper was to review the current state of the field of assessment and to speculate on what the future of assessments might be. We considered favorable assessment research directions based on our review. The future of assessments is largely about the future of education and work, and the skills we as a society will seek in the future, and thus we first considered what skills might be the most viable for the future. Our analysis of future skills was based on trend analyses, employer surveys, analyses of technology impacts, expert opinions, and the ETS (2023) survey of 17,000 adults in 17 middle and high-income countries. We next considered promising, innovative approaches for measuring those skills, focusing particularly on methods for measuring *hard-to-measure skills*. Next, we considered testing operations, from administration considerations through item development, including personalization, security, scoring, and evaluation, highlighting the role AI and technology will play in enhancing those operations. Finally, we considered the topic of feedback to the test-taker from the standpoint of learning science and based on demands from test-takers and other stakeholders.

There are several broad conclusions we draw based on our review of findings. First, advances in technology, particularly AI, will have profound effects, which we are only beginning to grasp, on all aspects of assessment, from what skills will be measured to how we will go about measuring them, how we will report results back to test-takers and stakeholders, and what we might expect recipients to do with those results.

Second, a core set of soft skills, durable skills, and complex skills are likely to become increasingly important in the future. The history of assessment, particularly the assessment of educational achievement and workplace skills, has largely focused on curricular and technical skills. Assessment of these skills will continue to be important, particularly the assessment of change and growth in these skills, but there is a newfound recognition that soft skills are as important if not more important for success in school, the workplace, and life. Social skills—teamwork, collaboration, communications—are likely to gain prominence based on occupational trends. Adaptability is likely to be increasingly important as AI and technology-driven changes will affect what workers will be asked to do throughout their careers, putting a premium on lifelong continuous learning, for fulfillment and well-being as well as for financial stability. Creativity and critical thinking will become increasingly important because these are skills for which humans hold an advantage over computers, which is likely to remain the case for some time, and these are skills that are likely to be augmented rather than replaced by AI.

Along with this increasing role placed on skills emerging over the lifespan will be a system in place to assess and recognize skill development. A strong majority of respondents around the world expressed the belief that non-degree credentials will become a valuable way to showcase skills and that in the future such proof of specific skills will become more important than a university degree. Such credentials may come from a university but will be treated as equally valuable if they come from a company or standardized testing or learning assessment organization. Relying on assessments to gain microcredentials and other certifications of skill acquisition will elevate the importance of security issues for those certifications.

A fourth conclusion is that we do not have good assessments for many of the skills that are likely to be increasingly important in the future. There is some skepticism about the quality of the measures that are available for many of these skills, which often rely on impressions, self-reports, and other subjective methods. This presents a tremendous opportunity for the field of assessment to develop rigorous, psychometrically sound assessments of skills that currently are considered hard to measure.

Finally, attitudes towards assessments are quite positive. Assessments motivate test users to acquire new skills and allow them to feel confident and prepared to pursue opportunities and advance careers, which will become increasingly important with AI-driven changes in the workplace. Assessments are seen by many as boosting self-esteem and career satisfaction and bridging the skills gap to provide equal opportunities for advancements across different backgrounds. This important role for assessments is contingent on the kind of feedback and insights about themselves that test-takers can gain by taking the assessment. Providing personalized, useful, actionable feedback to test-takers is an important and achievable goal for future assessment.

Limitations

There are limitations to our efforts to predict the future of assessments, like the limitations of predicting the future generally. People are not very accurate in predicting the future (Grossman, 2023; Rees, 2021.) However, as an organization partly responsible for designing some part of that future, ETS may have advantages over those who are merely forecasting it. Grossman (2023) suggested that better forecasters have “scientific expertise in a prediction domain, were interdisciplinary, used simpler models, and based predictions on prior data.” The report writing team along with the reviewers have assessment expertise from various perspectives and we leaned on expertise from the interviewees participating in the ETS (2023a) Human Progress Study. Through the participating external experts from the ETS Human Progress Study and on the report writing team, we did approach the task from an interdisciplinary perspective, and in reviewing the broad literature we relied on prior data. Rees (2021) suggested a crowd-sourcing strategy to overcome individual biases and we could argue that the ETS Human Progress Study, providing data from more than 17,000 respondents in 17 countries representing a variety of backgrounds provides that. Nevertheless, we did not employ a systematic forecasting methodology and our predictions about the future of assessments must therefore be interpreted with caution.

Another limitation is that we did not give all areas of assessment equal attention, opting instead to focus on areas we believed would experience the greatest changes due to technology, AI, learning science developments, and gaps between possibilities and current status, such as the opportunities to make assessment more useful to test-takers. Our focus in Section 2 was on the skills that are likely to grow in importance due to technology-driven changes. The skills and knowledge that facilitate mastery and application of foundational literacies, language, mathematics, and other school subjects that were addressed in National Research Council (2012) will remain important. The topics we covered in Sections 3 to 5, on new measurement methods, operations, and feedback apply to foundational literacies and other school subjects as well as to the new, durable skills that were the focus of Section 2. We did not focus on two important areas of K-12 assessment, classroom assessments, and accountability. One can imagine a companion report that would explore more deeply the relationships between the skills focused on for this report and the more traditional academic skills and content in reading, mathematics, and science with an emphasis on classroom assessments and accountability, perhaps extended internationally.

Future directions

There are several key research issues and directions we wish to suggest, which align with the main sections of the paper. First, the changing nature of skills needs monitoring—the skills demanded in the workforce affect educational standards and the curriculum down the line, and it is therefore useful to anticipate those changes. Second, richer assessment methods involving the characterization of learning within assessment, and exploring new, innovative approaches, including collaborative and multimodal approaches, will almost certainly gain increased research attention, and the OECD’s (2023) suggestions of considerations regarding innovative assessments of complex skills seem useful to pursue. Third, the various aspects of testing operations—item development, personalization, scoring, security, and reporting—are already being affected by rapid developments in technology and AI, and the pace of change in those operations is not likely to slow. Finally, almost every paper of the dozens on the future of assessments issued over the past decade has issued a plea or predicted some kind of advancement in providing useful, actionable feedback to test-takers to provide them with insights on where they are and how they can improve, and we endorse that plea.

ETS Research Institute is responding to this direction through four strands of research. These focus on personalizing assessment, designing principles for the creation of innovative, interactive digital assessments, developing standards for responsible and ethical AI applications including automated content generation and scoring, and impacting policy and practice through conceptualizing next-generation educational systems that close disparities. Through the research outlined here and through ETS Research Institute's research strands we may be positioned to repurpose assessment to better serve human learning, without relinquishing its traditional role in measuring achievement and developed ability. This will move us closer to a vision of assessment outlined in a collection of papers issued a decade ago by the Gordon Commission on The Future of Assessment in Education (2013).

Finally, to facilitate advances in education and skills assessment that will enable achievement of this vision, we call for a significant research investment. The global education expenditure is over $5 trillion per year, approximately 6% of global gross domestic product (World Economic Forum, 2022). Yet only a small part of that investment is concerned with assessment, which is needed to serve human learning and to monitor educational progress. The World Economic Forum's (2021) Global Taxonomy of Skills at Work provides a vision of a skills-based labor market; the companion World Economic Forum's (2023) Education 4.0 Framework identifies the *content skills* of global citizenship, innovation and creativity, technology skills, and interpersonal skills as critical for preparing the next generation for the future of work and societies. The latter three skills are well-aligned with those skills we identified in Section 2 of this report, based on our analyses of current and future workforce demands. Education 4.0 also identifies critical developments in learning experiences—personalized and self-paced, accessible and inclusive, problem-based and collaborative, lifelong and student-driven, which also are well-aligned with the themes we have identified throughout this report. Advances in assessment, which are achievable through focus and investment, will play a central role in moving us closer to achieving the visions articulated in the reports by the Gordon Commission (2013) and the World Economic Forum (2021; 2023).

References

AAMC PREview® professional readiness exam. (n.d.). *Students & residents*. Retrieved April 3, 2024, from https://students-residents.aamc.org/aamc-preview/aamc-preview-professional-readiness-exam

Abrams, Z. (2024). Addressing equity and ethics in artificial intelligence. *Monitor on Psychology*, *55*(3), 24–29. https://www.apa.org/monitor/2024/04/addressing-equity-ethics-artificial-intelligence

Abyaa, A., Khalidi Idrissi, M., & Bennani, S. (2019). Learner modelling: systematic review of the literature from the last 5 years. *Educational Technology Research and Development, 67*, 1105-1143.

Acar, S. (2023). Creativity assessment, research, and practice in the age of artificial intelligence. *Creativity Research Journal*, 1-7. Advance online publication. https://doi.org/10.1080/10400419.2023.2271749

Acuity Insights. (n.d.). *What is Casper?* https://acuityinsights.app/casper/

Acuity Insights. (2023). *Casper technical manual.* https://acuityinsights.com/casper-technical-manual/

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of educational research*, *87*(3), 659-701.

Agrawal, A., Gans, J., & Goldfarb, A. (2022). *Power and prediction: The disruptive economics of artificial intelligence.* Harvard Business Review Press.

Aguilar, S. J., Stuart A. Karabenick, S. A., Stephanie D. Teasley, S. D., Clare Baek, C. (2021). Associations between learning analytics dashboard exposure and motivation and self-regulated learning, *Computers & Education, 162*, 104085, https://doi.org/10.1016/j.compedu.2020.104085.

Ahn, T., Arcidiacono, P., Hopson, A., & Thomas, J. R. (2019). *Equilibrium grade inflation with implications for female interest in STEM majors* (Working Paper 26556). National Bureau of Economic Research. https://doi.org/10.3386/w26556

Alan, S., Boneva, T., & Ertac, S. (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics, 134*(3), 1121–1162. https://doi.org/10.1093/qje/qjz006

Ali, U. S., & van Rijn, P. W. (2016). An evaluation of different statistical targets for assembling parallel forms in item response theory. *Applied Psychological Measurement*, *40*(3), 163–179. https://doi.org/10.1177/0146621615613308

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

American Psychological Association. (2018). *Top 20 principles from psychology for preK-12 teaching and learning: Coalition for psychology in schools and education*. https://www.apa.org/ed/schools/teaching-learning/top-twenty-principles.pdf

Association of Test Publishers. (2022). *Guidelines for technology-based assessment*. https://www.testpublishers.org/assets/TBA%20Guidelines%203-14-2022%20draft%20numbered.pdf

Attali, Y., & van der Kleij, F. (2017). Effects of feedback elaboration and feedback timing during computer-based practice in mathematics problem solving. *Computers & Education*, 110, 154–169. https://doi.org/10.1016/j.compedu.2017.03.012

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & Von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5, Article 903077. https://doi.org/10.3389/frai.2022.903077

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, 118(4), 1279–1333. https://doi.org/10.1162/003355303322552801

Autor, D., Chin, C., Salomons, A., & Seegmiller, B. (2024). New frontiers: The origins and content of new work, 1940–2018. *The Quarterly Journal of Economics*. Advance online publication. https://doi.org/10.1093/qje/qjae008

Azevedo, R., & Bernard, R. M. (1995). The effects of computer-presented feedback on learning from computer-based instruction: A meta-analysis. *Journal of Educational Computing Research, 13*(2) 111–127. https://doi.org/10.2190/9LMD-3U28-3A0G-FTQT

Baker, R. S. J. d., & Yacef, K. (2009). *The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1*(1), 3–17. https://doi.org/10.5281/zenodo.3554657

Bailey, T., Jeong, D. W., & Cho, S. W. (2010). Referral, enrollment, and completion in developmental education sequences in community colleges. *Economics of Education Review, 29*(2), 255–270. https://doi.org/10.1016/j.econedurev.2009.09.002

Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of educational research, 61*(2), 213-238. https://doi.org/10.3102/00346543061002213

Bauer, M. S., Damschroder, L., Hagedorn, H., Smith, J., & Kilbourne, A. M. (2015). An introduction to implementation science for the non-specialist. *BMC Psychology, 3*(32), 1–12. https://doi.org/10.1002/j.2333-8504.2002.tb01890.x

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2002). *A feasibility study of on-the-fly item generation in adaptive testing* (Research Report No. RR-02-03). ETS. https://doi.org/10.1002/j.2333-8504.2002.tb01890.x

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1–27). Lawrence Erlbaum Associates.

Bennett, R. E. (1998). *Reinventing assessment:. speculations on the future of large-scale educational testing* (Policy Information Perspective). ETS. http://www.ets.org/Media/Research/pdf/PICREINVENT.pdf

Bennett, R. E. (2011). *Formative assessment: A critical review. Assessment in Education: Principles, Policy & Practice, 18*(1), 5-25. https://doi.org/10.1080/0969594X.2010.513678

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educational Assessment, 28*(2), 83–104. https://doi.org/10.1080/10627197.2023.2202312

Berman, A. I., Feuer, M. J., & Pellegrino, J. W. (2019). What use is educational assessment? *The Annals of the American Academy of Political and Social Science*, 683(1), 8–20. https://doi.org/10.1177/0002716219843871

Bernacki, M. L. (2018). Examining the cyclical, loosely sequenced, and contingent features of self-regulated learning: Trace data and their analysis. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 370–387). https://doi.org/10.4324/9781315697048-24

Biddle, D. A., & Nooren, P. M. (2006). Validity generalization vs. Title VII: Can employers successfully defend tests without conducting local validation studies? *Labor Law Journal*, 57, 216–237. https://testgenius.com/articles/validity-generalization.pdf

Bicknell, K., Brust, C., & Settles, B. (2023, February 5). How Duolingo's AI learns what you need to learn. *IEEE Spectrum*. https://spectrum.ieee.org/duolingo

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Black, P., & Wiliam, D. (1998). *Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Blackman, R., & Ammanath, B. (2022, March 21). Ethics and AI: 3 conversations companies need to have. *Harvard Business Review.* https://hbr.org/2022/03/ethics-and-ai-3-conversations-companies-need-to-be-having

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher, 13*(6), 4–16.

Bolsinova, M., Deonovic, B., Arieli-Attali, M., Burr, S., Hagiwara, M., & Maris, G. (2022). Measurement of ability in adaptive learning and assessment systems when learners use on-demand hints. *Applied Psychological Measurement*, 46(3), 219–235. https://doi.org/10.1177/01466216221084208

Bratsberg, B., & Rogeberg, O. (2018). Flynn effect and its reversal are both environmentally caused. *Proceedings of the National Academy of Sciences*, *115*(26), 6674–6678. https://doi.org/10.1073/pnas.1718793115

Bresnahan, T. (2010). General purpose technologies. In B. H. Hall & N. Rosenberg (Eds.), *Handbook of the economics of innovation* (Vol. 2, pp. 761–791). https://doi.org/10.1016/S0169-7218(10)02002-2

Brookhart, S., Stiggins, R., McTighe, J., & Wiliam, D. (2020). *The future of assessment practices: Comprehensive and balanced assessment systems. Learning Sciences International.* https://testing123.education.mn.gov/cs/groups/communications/documents/document/mdaw/mdaw/~edisp/000231.pdf

Bradley, M. (1975). Scientific education versus military training: *The influence of Napoleon Bonaparte on the Ecole Polytechnique. Annals of Science*, *32*(5), 415–449. https://doi.org/10.1080/00033797500200381

Bryk, A. S., & LeMahieu. (2015). *Learning to improve: How America's schools can get better at getting better*. Harvard Education Publishing. https://www.carnegiefoundation.org/resources/publications/learning-to-improve/

Buckley, J., Colosimo, L., Kantar, R., McCall, M., & Snow, E. (2021). *Game-based assessment for education. In OECD digital education outlook 2021: Pushing the frontiers with artificial intelligence, blockchain and robots* (pp. 195–208). OECD. https://read.oecd-ilibrary.org/education/oecd-digital-education-outlook-2021_9289cbfd-en#page1

Bull, S., & Kay, J. (2016). SMILI: A framework for interfaces to learning data in open learner models, learning analytics and related fields. *International Journal of Artificial Intelligence in Education*, 26, 293-331. https://doi.org/10.1007/s40593-015-0090-8

Burning Glass Technologies. (2019). *Mapping the genome of jobs: The Burning Glass skills taxonomy* [White paper]. https://www.voced.edu.au/content/ngv%3A84406

Burrus, J., Rikoon, S. H., & Brenneman, M. W. (Eds.). (2022). *Assessing competencies for social and emotional learning: Conceptualization, development, and applications*. Routledge. https://doi.org/10.4324/9781003102243

BusinessWire. (2024). *Carnegie learning wins 2024 EdTech award for MATHstream* [Press release]. https://www.businesswire.com/news/home/20240327088407/en/Carnegie-Learning-Wins-2024-EdTech-Award-for-MATHstream

Buyse, T., & Lievens, F. (2011). Situational judgment tests as a new tool for dental student selection. *Journal of Dental Education*, 75(6), 743–749. https://doi.org/10.1002/j.0022-0337.2011.75.6.tb05101.x

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

cApStAn & Halleux, B. (2019). *PISA 2021 translation and adaptation guidelines*. OECD. https://www.oecd.org/pisa/pisaproducts/PISA-2022-Translation-and-Adaptation-Guidelines.pdf

Cao, M., Drasgow, F., & Cho, S. (2015). Developing ideal intermediate personality items for the ideal point model. *Organizational Research Methods, 18*(2), 252–275. https://doi.org/10.1177/1094428114555993

Casner-Lotto, J., & Barrington, L. (2006). *Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century US workforce*. Partnership for 21st Century Skills.

Cattell, R. B. (1965). A biometrics invited paper. Factor analysis: An introduction to essentials I. The purpose and underlying models. *Biometrics, 21*(1), 190–215. https://doi.org/10.2307/2528364

Cattell, R. B., & Warburton, F. W. (1967). *Objective personality and motivation tests: A theoretical introduction and practical compendium*. University of Illinois Press.

Chakraborty, M., Tonmoy, T. I., Zaman, M., Gautam, S., Kumar, T., Sharma, K., Barman, N., Gupta, C., Jain, V., Chadha, A., Sheth, A., & Das, A. (2023). Counter Turing test (CT2): AI-generated text detection is not as easy as you may think—Introducing AI Detectability Index (ADI). In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 2206–2239). ACL. https://aclanthology.org/2023.emnlp-main.136/

Cengage. (2019, January 16). *New survey: demand for "uniquely human skills" increases even as technology and automation replace some jobs* [Press release]. https://www.cengagegroup.com/news/press-releases/2019/new-survey-demand-for-uniquely-human-skills-increases-even-as-technology-and-automation-replace-some-jobs/

Chamorro-Premuzic, T. (2021, May 26). The problem with job interviews*. Forbes*. https://www.forbes.com/sites/tomaspremuzic/2021/05/26/the-problem-with-job-interviews/?sh=4292b1224dee

Chan, S., Somasundaran, S., Ghosh, D., & Zhao, M. (2022). AGReE: A system for generating automated grammar reading exercises. In W. Che & E. Shutova (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 169–177). Association for Computational Linguistics. https://aclanthology.org/2022.emnlp-demos.17/

Charness, G., Gneezy, U., & Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior & Organization*, 149, 74–87. https://doi.org/10.1016/j.jebo.2018.02.024

Chen, L., Feng, G., Joe, J., Leong, C. W., Kitchen, C., & Lee, C. M. (2014). Towards automated assessment of public speaking skills using multimodal cues. In *ICMI '14: Proceedings of the 16th International Conference on Multimodal Interaction* (pp. 200–203). ACM. https://doi.org/10.1145/2663204.2663265

Cheng, K. H. C., Hui, C. H., & Cascio, W. F. (2017). Leniency bias in performance ratings: The Big-Five correlates. *Frontiers in Psychology, 8,* Article 521. https://doi.org/10.3389/fpsyg.2017.00521

Chernyshenko, O. S., Kankaraš, M., & Drasgow, F. (2018). *Social and emotional skills for student success and well-being: Conceptual framework for the OECD study on social and emotional skills* (OECD Education Working Paper No. 173). OECD. https://one.oecd.org/document/EDU/WKP(2018)9/En/pdf

Chetty, R., Deming, D. J., & Friedman, J. N. (2023). *Diversifying society's leaders? The determinants of causal effects of admission to highly selective private colleges* (Working Paper No. 31492). National Bureau of Economic Research. https://doi.org/10.3386/w31492

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Choi, I., Hao, J., Deane, P., & Zhang, M. (2021). *Benchmark keystroke biometrics accuracy from high-stakes writing tasks (Research Report* No. RR-21-15). ETS. https://doi.org/10.1002/ets2.12326

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, *63*(1), 83–117. https://doi.org/10.1111/j.1744-6570.2009.01163.x

Chopade, P., Edwards, D., Khan, S. M., Andrade, A., & Pu, S. (2019, November). CPSX: using AI-machine learning for mapping human-human interaction and measurement of CPS teamwork skills. In *2019 IEEE International Symposium on Technologies for Homeland Security (HST)* (pp. 1-6). IEEE. https://doi.org/10.1109/HST47167.2019.9032906

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220. https://doi.org/10.1037/h0026256

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

College Board. (2023, September 27). SAT suite: Everything you need to know about the Digital SAT. *College Board Blog.* https://blog.collegeboard.org/everything-you-need-know-about-digital-sat

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*(6), 1092–1122. https://doi.org/10.1037/a0021212

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90*(2), 218–244. https://doi.org/10.1037/0033-2909.90.2.218

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*, 253–278. https://doi.org/10.1007/BF01099821

Cotra, A. (2023, August 29). Language models surprised us. *Planned Obsolescence.* https://www.planned-obsolescence.org/language-models-surprised-us/

Cox, C. B., Barron, L. G., Davis, W., & de la Garza, B. (2017). Using situational judgment tests (SJTs) in training: Development and evaluation of a structured, low-fidelity scenario-based training method. *Personnel Review, 46*(1), 36–45. https://doi.org/10.1108/PR-05-2015-0137

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist, 30*(1), 1–14. https://doi.org/10.1037/0003-066X.30.1.1

Darling-Hammond, L. (2001). Inequality in teaching and schooling: How opportunity is rationed to students of color in America. In B. D. Smedley, A. Y. Stith, L. Colburn, & C. H. Evans (Eds.), *The right thing to do, the smart thing to do: Enhancing diversity in health professions—Summary of the Symposium on Diversity in Health Professions in Honor of Herbert W. Nickens, M. D.* (pp. 208–233). National Academies Press. http://www.nap.edu/catalog/10186.html

Davey, T. (2023). Automated test assembly. In R. J. Tierney, F. Rizvi, & K. Ercikan (Eds.), *International encyclopedia of education: Vol. 14. Quantitative research and educational measurement* (pp. 201–208). Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.10027-2

Davoli, M., & Entorf, H. (2018). *The PISA shock, socioeconomic inequality, and school reforms in Germany* (IZA Policy Paper No. 140). IZA – Institute of Labor Economics. https://docs.iza.org/pp140.pdf

De Boeck, P. (2023, July 25–28). *Pervasive DIF and DIF detection bias* [Paper presentation]. International Meeting of the Psychometric Society (IMPS 2023), University of Maryland, College Park, MD, United States.

De Boeck, P., & Cho, S.-J. (2021). Not all DIF is shaped similarly. *Psychometrika, 86*(3), 712–716. https://doi.org/10.1007/s11336-021-09772-3

Dell. (2018, January 30). *3,800 business leaders declare: It's A tale of two futures*. https://www.dell.com/en-us/perspectives/3800-business-leaders-declare-its-a-tale-of-two-futures/

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics, 132*(4), 1593–1640. https://doi.org/10.1093/qje/qjx022

Deming, D. (2024, March 7). The worst way to do college admissions: Making standardized test scores optional has harmed the disadvantaged applicants it was intended to help. *The Atlantic*. https://theatlantic.com/ideas/archive/2024/03/standardized-testing-requirements-act-sat/677667/

Deonovic, B., Yudelson, M., Bolsinova, M., Attali, M., & Maris, G. (2018). Learning meets assessment. *Behaviormetrika, 45*(2), 457–474. https://doi.org/10.1007/s41237-018-0070-z

Diao, Q., & van der Linden, W. J. (2013). Integrating test-form formatting into automated test assembly. *Applied Psychological Measurement, 37*(5), 361-374. https://doi.org/10.1177/0146621613476157

Di Battista, A., Grayling, S., Hasselaar, E., Leopold, T., Li, R., Rayner, M., & Zahidi, S. (2023, May). *Future of jobs report 2023.* World Economic Forum. https://www.weforum.org/reports/the-future-of-jobs-report-2023

DiCerbo, K. (2024, March 7). How we built AI tutoring tools. *Khan Academy Blog*. https://blog.khanacademy.org/how-we-built-ai-tutoring-tools/

Dobrescu, L., Holden, R., Motta, A., Piccoli A., Roberts, P., & Walker, S. (2021). *Cultural context in standardized tests* (Working Paper 2021-08). University of New South Wales Business School. https://doi.org/10.2139/ssrn.3983663

Duolingo Team. (2023, March 14). Introducing Duolingo Max, a learning experience powered by GPT-4. *Duolingo Blog*. https://blog.duolingo.com/duolingo-max/

Eberly Center. (n.d.). *Learning principles: Theory and research-based principles of learning.* Carnegie Mellon University. https://www.cmu.edu/teaching/principles/learning.html

Elliott, S. W. (2017). *Computers and the future of skill demand*. OECD. https://doi.org/10.1787/9789264284395-en

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs: An early look at the labor market impact potential of large language models*. arXiv. https://arxiv.org/abs/2303.10130v4

Embretson, S. (1994). Applications of cognitive design systems to test development. In C. R. Reynolds (Ed), *Cognitive assessment: A multidisciplinary perspective* (pp. 107 -135). Springer. https://doi.org/10.1007/978-1-4757-9730-5_6

Emerson, A., Houghton, P., Chen, K., Basheerabad, V., Ubale, R., & Leong, C. W. (2022). Predicting user confidence in video recordings with spatio-temporal multimodal analytics. In *ICMI '22 companion: Companion publication of the 2022 International Conference on Multimodal Interaction* (pp. 98-104). ACM. https://doi.org/10.1145/3536220.3558007

Erwin, T. D., & Sebrell, K. W. (2003). Assessment of critical thinking: ETS's tasks in critical thinking. *Journal of General Education, 52*(1), 50–70. https://doi.org/10.1353/jge.2003.0019

ETS. (n.d.). *Demonstrate program effectiveness with the ETS® Major Field Tests*. https://www.ets.org/mft.html

ETS. (2014). *ETS standards for quality and fairness*. https://ets.org/pdfs/about/standards-quality-fairness.pdf

ETS. (2022). *ETS guidelines for developing fair tests and communications*. https://www.ets.org/pdfs/about/fair-tests-and-communications.pdf

ETS. (2023a). *ETS human progress study. [Unpublished data set].*

ETS. (2023b). *Your at home testing*. https://www.ets.org/gre/test-takers/general-test/register/at-home-testing.html

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics, 133*(4), 1645–1692. https://doi.org/10.1093/qje/qjy013

Feuer, M. J. (2012). *No country left behind: Rhetoric and reality of international large-scale assessment.* ETS. http://www.ets.org/Media/Research/pdf/PICANG13.pdf

Feuer, M., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (Eds.). (1999). *Uncommon measures: Equivalence and linkage among educational tests*. National Academies Press. https://doi.org/10.17226/6332

Flanagan, C. (2021, July 22). The University of California is lying to us. *The Atlantic.* https://www.theatlantic.com/ideas/archive/2021/07/why-university-california-dropping-sat/619522/

Flynn, M. (2023, May 30). The soft skills “debate” is over. *Forbes*. https://www.forbes.com/sites/mariaflynn/2023/05/30/the-soft-skills-debate-is-over/?sh=5baa274b7308

Foster, N., & Piacentini, M. (Eds.). (2023). *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Frensch, P. A., & Funke, J. (1995). *Complex problem solving: The European perspective.* Routledge.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerization? *Technological Forecasting and Social Change, 114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Friedland, N. S., Allen, P. G., Matthews, G., Witbrock, M., Baxter, D., Curtis, J., Shepard, B., Miraglia, P., Angele, J., Staab, S., Moench, E., Oppermann, H., Wenke, D., Israel, D., Chaudhri, V., Porter, B., Barker, K., Fan, J., Chaw, S., … Clark, P. (2004). Project Halo: Towards a digital Aristotle. *AI Magazine, 25*(4), 29–47. https://doi.org/10.1609/aimag.v25i4.1783

Fu, J., Tan, A., & Kyllonen, P. C. (in press). The Rank-2PL IRT models for forced-choice questionnaires: Maximum marginal likelihood estimation with an EM algorithm. *Journal of Educational and Behavioral Statistics*.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional children, 53*(3), 199-208. Exceptional Children, *53*(3), 199 -208. https://doi.org/10.1177/001440298605300301

Fyfe, E. R., Borriello, G. A., & Merrick, M. (2023). A developmental perspective on feedback: How corrective feedback influences children’s literacy, mathematics, and problem solving. *Educational Psychologist, 58*(3), 130–145. https://doi.org/10.1080/00461520.2022.2108426

Fyfe, E. R., De Leeuw, J. R., Carvalho, P. F., Goldstone, R. L., Sherman, J., Admiraal, D., Alford, L. K., Bonner, A., Brassil, C. E., Brooks, C. A., Carbonetto, T., Chang, S. H., Cruz, L., Czymoniewicz-Klippel, Daniel, F., Driessen, M., Habashy, N., Hanson-Bradley, C. L., Hirt, E. R., … Motz, B. A. (2021). Many Classes 1: Assessing the generalizable effect of immediate feedback versus delayed feedback across many college classes. *Advances in Methods and Practices in Psychological Science*, *4*(3), Article 25152459211027575. https://doi.org/10.1177/25152459211027575

Gao, L., Ghosh, D., & Gimpel, K. (2022). What makes a question inquisitive? A study on type-controlled inquisitive question generation. In V. Nastase, E. Pavlick, M. T. Pilehvar, J. Camacho-Callados, & A. Raganato (Eds.), *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics* (pp. 240–257). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.starsem-1.22

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2414–2423). IEEE. https://doi.org/10.1109/CVPR.2016.265

Geerlings, H., Glas, C. A., & Van Der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika*, 76, 337-359. https://doi.org/10.1007/s11336-011-9204-x

Geiger, M., Bärwaldt, R., & Wilhelm, O. (2021). The good, the bad, and the clever: Faking ability as a socio-emotional ability? *Journal of Intelligence*, *9*(1), 1–22. https://doi.org/10.3390/jintelligence9010013

Gierl, M. J., & Haladyna, T. M. (Eds.). (2012). *Automatic item generation: Theory and practice*. Routledge.

Gil, Y., & Selman, B. (2019). *A 20-year community roadmap for artificial intelligence research in the US.* arXiv. https://doi.org/10.48550/arXiv.1908.02624

Glas, C. A. W., & van der Linden, W. J. (2001, June 2–4). *Modeling variability in item parameters in CAT* [Paper presentation]. North American Psychometric Society Meeting, King of Prussia, PA, United States.

Godwin, K. E., Almeda, M. V., Seltman, H., Kai, S., Skerbetz, M. D., Baker, R. S., & Fisher, A. V. (2016). Off-task behavior in elementary school children. *Learning and Instruction*, 44, 128–143. https://doi.org/10.1016/j.learninstruc.2016.04.003

Goldberg, B., & Sinatra, A. M. (2023). Generalized intelligent framework for tutoring (gift) SWOT analysis. In A. M. Sinatra, A. C. Graesser, X. Hu, G. Goodwin, & V. Rus (Eds.), *Design recommendations for intelligent tutoring systems: Vol.10. Strengths, weaknesses, opportunities and threats (SWOT) analysis of intelligent tutoring systems* (pp. 9–26). U.S. Army Combat Capabilities Development Command—Soldier Center. https://gifttutoring.org/documents/163

Goodhart, C. A. E. (1984). *Monetary theory and practice: The U.K. experience*. Springer. https://doi.org/10.1007/978-1-349-17295-5

The Gordon Commission on The Future of Assessment in Education. (2013). *To assess, to teach, to learn: A vision for the future of assessment* (Technical report). ETS. https://www.ets.org/Media/Research/pdf/gordon_commission_technical_report.pdf

Gosling, S. D., Augustine, A. A., Vazire, S., Holtzman, N., & Gaddis, S. (2011). Manifestations of personality in online social networks: Self-reported Facebook-related behaviors and observable profile information. *Cyberpsychology, Behavior, and Social Networking, 14*(9), 483–488. https://doi.org/10.1089/cyber.2010.0087

Graf, E. A., & Fife, J. H. (2012). Difficulty modeling and automatic generation of quantitative items: Recent advances and possible next steps. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation* (pp. 157-178). Routledge.

Greiff, S., Gašević, D., & von Davier, A. (2017). Using process data for assessment in intelligent tutoring systems: A cognitive psychologist, psychometrician, and computer scientist perspective. In R. Sottilare, A. Graesser, X. Hu, & G. Goodwin (Eds.), *Design recommendations for intelligent tutoring systems: Vol. 5. Assessment methods* (pp. 171–179). U.S. Army Research Laboratory. https://gifttutoring.org/attachments/download/2410/Design%20Recommendations%20for%20ITS_Volume%205%20-%20Assessment_final_errata%20corrected.pdf

Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin*, 124(1), 75–111. https://doi.org/10.1037/0033-2909.124.1.75

Grose, J. (2024, January 17). Don't ditch standardized tests: Fix them. *The New York Times*. https://www.nytimes.com/2024/01/17/opinion/standardized-tests.html

Grossmann, I. (2023). Insights into the accuracy of social scientists' forecasts of societal change. *Nature Human Behavior*, 7, 484–501. https://doi.org/10.1038/s41562-022-01517-1

Haberman, S. J., & Lee, Y.-H. (2017). *A statistical procedure for testing unusually frequent exactly matching responses and nearly matching responses* (Research Report No. RR-17-23). ETS. https://doi.org/10.1002/ets2.12150

Haberman, S. J., Lee, Y.-H., Papierman, P., Zhou, Y., & Subhedar, R. (2022). *Systems and methods for detecting unusually frequent exactly matching and nearly matching test responses* (U.S. Patent 11,398,161). U.S. Patent Office and Trademark Office. https://ppubs.uspto.gov/pubwebapp/external.html?q=(11398161).pn.&db=USPAT&type=ids

Hambleton, R. K. (2002). Adapting achievement tests into multiple languages for international assessments. In National Research Council (Ed), *Methodological advances in cross-national surveys of educational achievement* (pp. 58–79). National Academies Press. https://nap.nationalacademies.org/read/10322/chapter/4

Hao, J., von Davier, A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. (in press). Transforming assessment: the impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practices.*

Hattie, J. A. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hattie, J., & Gan, M. (2011). Instruction based on feedback. In E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction (pp. 249–271)*. Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hao, J., Liu, L., von Davier, A. A., Lederer, N., Zapata-Rivera, D., Jakl, P., & Bakkenson, M. (2017). *EPCAL: ETS platform for collaborative assessment and learning* (Research Report No. RR-17-49). ETS. https://doi.org/10.1002/ets2.12181

He, J., Bartram, D., Inceoglu, I., & van de Vijver, F. J. R. (2014). Response styles and personality traits: A multilevel analysis. *Journal of Cross-Cultural Psychology, 45*(7), 1028–1045. https://doi.org/10.1177/0022022114534773

He, Q., Borgonovi, F., & Paccagnella, M. (2019). *Using process data to understand adults' problem-solving behaviour in the programme for the international assessment of adult competencies (PIAAC): Identifying generalised patterns across multiple tasks with sequence mining* (OECD Educaton working paper No. 205). OECD. https://one.oecd.org/document/EDU/WKP(2019)13/en/pdf

Heckman, J., & Zhou, J. (2021). *Interactions as investments: The microdynamics and measurement of early childhood learning* [Unpublished manuscript]. Center for the Economics of Human Development and Department of Economics, University of Chicago.

Hedlund, J., Wilt, J. M., Nebel, K. L., Ashford, S. J., & Sternberg, R. J. (2006). Assessing practical intelligence in business school admissions: A supplement to the graduate management admissions test. *Learning and Individual Differences, 16*(2), 101–127. https://doi.org/10.1016/j.lindif.2005.07.005

Herman, J. L., Martínez, J. F., & Bailey, A. L. (2023). Fairness in educational assessment and the next edition of the standards: *Concluding commentary. Educational Assessment*, *28*(2), 128–136. https://doi.org/10.1080/10627197.2023.2215980

Hilton, M., & Herman, J. (Eds.). (2017). *Supporting students' college success: The role of assessment of intrapersonal and interpersonal competencies.* National Academies Press.

Himelfarb, I. (2019). A primer on standardized testing: History, measurement, classical test theory, item response theory, and equating. *Journal of Chiropractic Education*, *33*(2), 151–163. https://doi.org/10.7899/JCE-18-22

Hinnant-Crawford, B. N. (2020). *Improvement science in education: A primer.* Myers Education Press.

Hitt, C., Trivitt, J., & Cheng, A. (2016). When you say nothing at all: The predictive power of student effort on surveys. *Economics of Education Review,* 52, 105–119. https://doi.org/10.1016/j.econedurev.2016.02.001

Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-index: Statistical theory and empirical support* (Research Report No. RR-96-07). ETS. https://doi.org/10.1002/j.2333-8504.1996.tb01685.x

Hood, S. (1998). Culturally responsive performance-based assessment: Conceptual and psychometric considerations. *Journal of Negro Education*, *67*(3), 187–196. https://doi.org/10.2307/2668188

Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, *4*(4), 403–424. https://doi.org/10.1037/1082-989X.4.4.403

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., & Xing, E. P. (2017). Toward controlled generation of text. In D. Precup & Y. W. Teh (Eds.), *Proceedings of machine learning research: Vol. 70. Proceedings of the 34th International Conference on Machine Learning* (pp. 1587–1598). https://proceedings.mlr.press/v70/hu17e.html

IMS Global. (2022). *Question & test interoperability (QTI) 3.0: Best practices and implementation guide*. https://www.imsglobal.org/spec/qti/v3p0/impl/

Institute of Medicine. (2015). *Psychological testing in the service of disability determination*. The National Academies Press. https://doi.org/10.17226/21704

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing*, *1*(2), 93–114. https://doi.org/10.1207/S15327574IJT0102_1

International Test Commission. (2013). *ITC guidelines for test use.* Final version. https://www.intestcom.org/files/guideline_test_use.pdf

International Test Commission. (2017). *The ITC guidelines for translating and adapting tests* (2nd ed.). https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

International Test Commission & Association of Test Publishers. (2022). *Guidelines for technology-based assessment*. https://www.intestcom.org/upload/media-library/guidelines-for-technology-based-assessment-v20221108-16684036687NAG8.pdf

Irvine, S. H., & Kyllonen, P. C. (Eds.). (2013). *Item generation for test development*. Routledge.

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non-test score outcomes. *Journal of Political Economy, 126*(5), 2072–2107. https://doi.org/10.1086/699018

Jiang, Y., Martin-Raugh, M., Yang, Z., Hao, J., Liu, L., & Kyllonen, P. C. (2023). Do you know your partner's personality through virtual collaboration or egotiation? Investigating perceptions of personality and their impacts on performance. *Computers in Human Behavior*, 141, Article 107608. https://doi.org/10.1016/j.chb.2022.107608

Jimenez, L., & Modaffari, J. (2021). *Future of testing in education: Effective and equitable assessment systems.* Center for American Progress. https://www.americanprogress.org/article/future-testing-education-effective-equitable-assessment-systems/

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (Vol. 2., pp. 102–138). Guilford Press

Johnson, M. S. (2024). *How do we demonstrate AI responsibility: The devil is in the details*. [Manuscript in preparation].

Johnson, M. S., Liu, X., & McCaffrey, D. F. (2022). Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. *Journal of Educational Measurement, 59*(3), 338–361. https://doi.org/10.1111/jedm.12335

Johnson, M. S., & McCaffrey, D. F. (2023). Evaluating fairness of automated scoring in educational measurement. In S. Lane (Ed.), *Advancing natural language processing in educational assessment* (pp. 143–164) . Routledge. https://doi.org/10.4324/9781003278658-12

Johnson, M. S., & Sinharay, S. (2005). Calibration of polytomous item families using Bayesian hierarchical modeling. *Applied Psychological Measurement*, *29*(5), 369-400.

Jung, J. Y., Tyack, L., & von Davier, M. (2022). Automated scoring of constructed-response items using artificial neural networks in international large-scale assessment. *Psychological Test and Assessment Modeling*, *64*(4), 471-494.

Karay, Y., Reiss, B., & Schauber, S. K. (2020). Progress testing anytime and anywhere: Does a mobile-learning approach enhance the utility of a large-scale formative assessment tool? *Medical Teacher*, *42*(10), 1154–1162. https://doi.org/10.1080/0142159X.2020.1798910

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 331(6018), 772-775. https://doi.org/10.1126/science.1199327

Kautz, T., & Zanoni, W. (2014). *Measuring and fostering non-cognitive skills in adolescence: Evidence from Chicago Public Schools and the OneGoal Program*. University of Chicago.

Kell, H. J., Martin-Raugh, M. P., Carney, L. M., Inglese, P. A., Chen, L., & Feng, G. (2017). *Exploring methods for developing behaviorally anchored rating scales for evaluating structured interview performance* (Research Report No. RR-17-28). ETS. https://doi.org/10.1002/ets2.12152

Kessler, J. B., Low, C., & Sullivan, C. D. (2019). Incentivized resume rating: Eliciting employer preferences without deception. *American Economic Review, 109*(11), 3713–3744. https://doi.org/10.1257/aer.20181714

Khan, S. (2023, March 14). Harnessing GPT-4 so that all students benefit. A nonprofit approach for equal access. *Khan Academy Blog*. https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis*, *15*(1), 46–66. https://doi.org/10.1093/pan/mpl011

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational measurement: Issues and practice*, *30*(4), 28-37. https://doi.org/10.1111/j.1745-3992.2011.00220.x

Klieger, D. M., Kell, H. J., Rikoon, S., Burkander, K. N., Bochenek, J. L., & Shore, J. R. (2018). *Development of the behaviorally anchored rating scales for the skills demonstration and progression guide* (Research Report No. RR-18-24). ETS. https://doi.org/10.1002/ets2.12210

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Klute, M., Apthorp, H., Harlacher, J., & Reale, M. (2017). *Formative assessment and elementary school student academic achievement: A review of the evidence* (Report No. REL 2017-259). Regional Educational Laboratory Central.

Koedinger, K. R., Carvalho, P. F., Liu, R., & McLaughlin, E. A. (2023). *An astonishing regularity in student learning rate. Proceedings of the National Academy of Sciences, 120*(13), Article e2221311120. https://doi.org/10.1073/pnas.2221311120

Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D., & Graepel, T. (2014). Manifestations of user personality in website choice and behaviour on online social networks. *Machine Learning*, 95, 357–380. https://doi.org/10.1007/s10994-013-5415-y

Krachman, S. B., Arnold, R., & LaRocca, R. (2016). *Expanding the definition of student success: A case study of the CORE districts.* Transforming Education. https://transformingeducation.org/wp-content/uploads/2017/04/TransformingEducationCaseStudyFINAL1.pdf

Kukea Shultz, P., & Englert, K. (2021). Cultural validity as foundational to assessment development: An indigenous example. *Frontiers in Education, 6,* Article 701973. https://doi.org/10.3389/feduc.2021.701973

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of educational research,* 86(1), 42-78. https://doi.org/10.3102/0034654315581420

Kumar, V., & Boulanger, D. (2020). *Explainable automated essay scoring: Deep learning really has pedagogical value. Frontiers in Education, 5*, Article 572367 https://doi.org/10.3389/feduc.2020.572367

Kuncel, N. R., Kochevar, R. J., & Ones, D. S. (2014). A meta-analysis of letters of recommendation in college and graduate admissions: Reasons for hope. *International Journal of Selection and Assessment, 22*(1), 101–107. https://doi.org/10.1111/ijsa.12060

Kyllonen, P. C., & Bertling, J. P. (2013). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 277–285). CRC Press.

Kyllonen, P. C. (2016). Socio-emotional and self-management variables in learning and assessment. In A. A. Rupp & J. P. Leighton (Eds.), *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 174–197). John Wiley & Sons. https://doi.org/10.1002/9781118956588.ch8

Kyllonen, P., Hartman, R., Sprenger, A., Weeks, J., Bertling, M., McGrew, K., Kriz, S., Bertling, J., Fife, J., & Stankov, L. (2019). General fluid/inductive reasoning battery for a high-ability population. *Behavior Research Methods*, *51*(2), 507–522. https://doi.org/10.3758/s13428-018-1098-4

Kyllonen, P. (2021). Taxonomy of cognitive abilities and measures for assessing artificial intelligence and robotics capabilities. In *AI and the future of skills: Vol. 1: Capabilities and assessments* (pp. 50–76).OECD Publishing. https://doi.org/10.1787/feecd512-en

Kyllonen, P., Hao, J., Weeks, J., Fauss, M., & Kerzabi, E. (2023). *Collaborative problem solving (CPS) skill: Estimating an individual's contribution to small group performance* [Unpublished manuscript]. ETS.

Kyriazos, T. A. (2018). Applied psychometrics: The application of CFA to multitrait-multimethod matrices (CFA-MTMM). *Psychology*, *9*(12), 2625–2648. https://doi.org/10.4236/psych.2018.912150

Landers, R. N., Armstrong, M. B., Collmus, A. B., Mujcic, S., & Blaik, J. (2022). Theory-driven game-based assessment of general cognitive ability: Design theory, measurement, prediction of performance, and test fairness. *Journal of Applied Psychology, 107*(10), 1655–1677. https://doi.org/10.1037/apl0000954

Landers, R. N., & Sanchez, D. R. (2022). Game-based, gamified, and gamefully designed assessments for employee selection: Definitions, distinctions, design, and validation. *International Journal of Selection and Assessment, 30*(1), 1–13. https://doi.org/10.1111/ijsa.12376

Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2016). *Handbook of test development* (Vol. 2, pp. 3–18). Routledge.

Lang, J. W. B., & Tay, L. (2021). The science and practice of item response theory in organizations. *Annual Review of Organizational Psychology and Organizational Behavior,* 8, 311–338. https://doi.org/10.1146/annurev-orgpsych-012420-061705

Langer, C., & Wiederhold, S. (2023). *The value of early-career skills* (CESifo Working Paper No. 10288). CESifo Network. https://doi.org/10.2139/ssrn.4369987

Lassébie, J., & Quintini, G. (2022). *What skills and abilities can automation technologies replicate and what does it mean for workers? New evidence* (OECD Social, Employment and Migration Working Papers, No. 282). OECD Publishing. https://doi.org/10.1787/646aad77-en

Law, K. S., Mobley, W. H., & Wong, C.-S. (2002). Impression management and faking in biodata scores among Chinese job-seekers. *Asia Pacific Journal of Management*, 19, 541–556. https://doi.org/10.1023/A:1020521726390

Lederman, O., Calacci, D., MacMullen, A., Fehder, D. C., Murray, F. E., & Pentland, A. S. (2016). Open badges: A low-cost toolkit for measuring team communication and dynamics. In *The online proceedings of the 2016 International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation* (SBP-BriMS 2016). http://sbp-brims.org/2016/proceedings/IN_105.pdf

Lee, G. H., Lee, K. J., Jeong, B. & Kim, T. (2024). Developing personalized marketing service using generative AI. *IEEE Access*, 12, 22394–22402. https://doi.org/10.1109/ACCESS.2024.3361946

Lee, Y.-H., & Haberman, S. J. (2013). Harmonic regression and scale stability. *Psychometrika*, *78*(4), 815–829. https://doi.org/10.1007/s11336-013-9337-1

Lee, Y.-H., & Haberman, S. J. (2021). Studying score stability with a harmonic regression family: A comparison of three approaches to adjustment of examinee-specific demographic data. *Journal of Educational Measurement*, *58*(1), 54–82. https://doi.org/10.1111/jedm.12266

Lee, Y.-H., & Lewis, C. (2021). Monitoring item performance with CUSUM statistics in continuous testing. *Journal of Educational and Behavioral Statistics, 46*(5), 611–648. https://doi.org/10.3102/1076998621994563

Lee, Y.-H., Lewis, C., & von Davier, A. A. (2014). Monitoring the quality and security of multistage tests. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 285–300). CRC Press.

Lee, Y.-H., & von Davier, A. A. (2013). Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. *Psychometrika*, *78*(3), 557–575. https://doi.org/10.1007/s11336-013-9317-5

Leenknecht, M., Hompus, P., & van der Schaaf, M. (2019). Feedback seeking behaviour in higher education: The association with students' goal orientation and deep learning approach. *Assessment & Evaluation in Higher Education, 44*(7), 1069–1078. https://doi.org/10.1080/02602938.2019.1571161

Lehman, B., Sparks, J. R., & Zapata-Rivera, D. (2018). When should an adaptive assessment care? In N. Guin & A. Kumar (Eds.), P*roceedings of ITS 2018: Intelligent Tutoring Systems 14th International Conference, Workshop on Exploring Opportunities for Caring Assessments* (pp. 87–94). ITS. https://ceur-ws.org/Vol-2354/w3paper1.pdf

Leonhardt, D. (2023, January 7). The misguided war on the SAT: Colleges have fled standardized tests, on the theory that they hurt diversity. That's not what the research shows. *The New York Times.* https://www.nytimes.com/2024/01/07/briefing/the-misguided-war-on-the-sat.html

Lewis, C. (2001). Expected response functions. In A. Boomsma, M. A J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 163–171). Springer. https://doi.org/10.1007/978-1-4613-0169-1_9

LinkedIn Talent Solutions. (2019). *Global talent trends: The 3 trends transforming your workplace*. https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/resources/pdfs/global_talent_trends_2019_emea.pdf

Lira, B., O'Brien, J. M., Peña, P. A., Galla, B. M., D'Mello, S., Yeager, D. S., Defnet, A., Kautz, T., Munkacsy, K., & Duckworth, A. L. (2022). Large studies reveal how reference bias limits policy applications of self-report measures. *Scientific Reports*, 12, Article 19189. https://doi.org/10.1038/s41598-022-23373-9

Lissitz, R. W. (2009). Introduction. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 1–15). IAP Information Age Publishing.

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher*, *41*(9), 352–362. https://doi.org/10.3102/0013189X12459679

Liu, O. L., Kell, H. J., Liu, L., Ling, G., Wang, Y., Wylie, C., Sevak, A., Sherer, D., LeMahieu, P., & Knowles, T. (2023). *A new vision for skills-based assessment*. ETS. https://ets.org/pdfs/rd/new-vision-skills-based-assessment.pdf

Liu, O. L., Mao, L., Frankel, L., & Xu, J. (2016). Assessing critical thinking in higher education: The HEIghten approach and preliminary validity evidence. *Assessment & Evaluation in Higher Education, 41*(5), 677–694. https://doi.org/10.1080/02602938.2016.1168358

Liu, X., Zhang, Z., Wang, Y., Pu, H., Lan, Y., & Shen, C. (2023). COCO: Coherence-enhanced machine-generated text detection under low resource with contrastive learning. In H. Bouamor, J. Pino, & K. Bali (Eds.), Proceedings of the 2023 *Conference on Empirical Methods in Natural Language Processing* (pp. 16167–16188). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.1005

Loewus, L. (2016, November 8). What is digital literacy? *Education Week.* https://www.edweek.org/teaching-learning/what-is-digital-literacy/2016/11

Ludlow, L. H., O'Keefe, T., Braun, H., Anghel, E., Szendey, O., Matz, C., & Howell, B., (2022). An enhancement to the theory and measurement of purpose. *Practical Assessment, Research, and Evaluation 27*(1), Article 4. https://doi.org/10.7275/c5jb-rr95

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. J*ournal of Educational Psychology, 106*(4), 901 -918. https://doi.org/10.1037/a0037123

MacCann, C., & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: Theory and data. *Emotion*, 8(4), 540–551. https://doi.org/10.1037/a0012746

Madnani, N., & Cahill, A. (2018). Automated scoring: Beyond natural language processing. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1099–1109). Association for Computational Linguistics. https://aclanthology.org/C18-1094

Mammadov, S. (2022). Big Five personality traits and academic performance: A meta- analysis. *Journal of Personality*, *90*(2), 222-255. https://doi.org/10.1111/jopy.12663

Mankki, V. (2023). Research using teacher or teacher educator job advertisements: A scoping review. *Cogent Educatio*n, *10*(1), Article 2223814. https://doi.org/10.1080/2331186X.2023.2223814

Martin-Raugh, M. P., Kyllonen, P. C., Hao, J., Bacall, A., Becker, D., Kurzum, C., Yang, Z., Yan, F., & Barnwell, P. (2020). Negotiation as an interpersonal skill: Generalizability of negotiation outcomes and tactics across contexts at the individual and collective levels. *Computers in Human Behavior*, 104, Article 105966. https://doi.org/10.1016/j.chb.2019.03.030

Martín-Raugh, M., Roohr, K. C., Leong, C. W., Molloy, H., McCulla, L., Ramanarayan, V., & Mladineo, Z. (2023). Better understanding oral communication skills: The impact of perceived personality traits. *American Journal of Distance Education*. Advance online publication. https://doi.org/10.1080/08923647.2023.2235950

Mattingly, S. M., Gregg, J. M., Audia, P., Bayraktaroglu, A. E., Campbell, A. T., Chawla, N. V., Das Swain, V., De Choudhury,M., D'Mello, S. K., Dey, A. K., Gao, G., Jagannath, K., Jiang, K., Lin, S., Liu, Q., Mark, G., Martinez, G. J. Masaba, K., Mirjafari, S., ... Striege, A. (2019, May). The tesserae project: Large-scale, longitudinal, in situ, multimodal sensing of information workers. In *Extended abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-8). ACM. https://doi.org/10.1145/3290607.3299041

McLaughlin, K., Ainslie. M., Coderre, S., Wright, B., & Violato, C. (2009). The effect of differential rater function over time (DRIFT) on objective structured clinical examination ratings. *Medical Education*, *43*(10), 989–992. https://10.1111/j.1365-2923.2009.03438.x

McWhorter, J. (2024, March 14). No, the SAT isn't racist. *The New York Times*. https://www.nytimes.com/2024/03/14/opinion/sat-college-admissions-antiracism.html

Mervosh, S. (2022, September 1). The pandemic erased two decades of progress in math and reading: The results of a national test showed just how devastating the last two years have been for 9-year-old schoolchildren, especially the most vulnerable. *The New York Times*. https://www.nytimes.com/2022/09/01/us/national-test-scores-math-reading-pandemic.html

Meyer, R. H., Wang, C., & Rice, A. B. (2018). *Measuring students' social-emotional learning among California's CORE districts: An IRT modeling approach* [Working paper]. Policy Analysis for California Education. https://edpolicyinca.org/sites/default/files/Measuring_SEL_May-2018.pdf

Mignogna, G., Carey, C. E., Wedow, R., Baya, N., Cordioli, M., Pirastu, N., Bellocco, R., Mlerbi, K. F., Nivard, M. G., Neale, B. M., Walters, R. K., & Ganna, A. (2023). Patterns of item nonresponse behaviour to survey questionnaires are systematic and associated with genetic loci. *Nature Human Behavior, 7*, 1371–1387. https://doi.org/10.1038/s41562-023-01632-7

Millsap, R. (2011). *Statistical approaches to measurement invariance*. Routledge.

Mirjafari, S., Masaba, K., Grover, T., Wang, W., Audia, P., Campbell, A. T., Chawla, N. V., Das Swain, V., De Choudhury, M., Dey, A. K., D'Mello, S. K., Gao, G., Gregg, J. M., Jagannath, K., Jiang, K., Lin, S., Qiang, L., Mark, G., Martinez, G. J., Martinez, S. M., .. Striegel, A. (2019). Differentiating higher and lower job performers in the workplace using mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *3*(2), 1–24. https://doi.org/10.1145/3328908

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement*, *30*(1), 55–78. https://doi.org/10.1111/j.1745-3984.1993.tb00422.x

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3 -62. https://doi.org/10.1207/S15366359MEA0101_02

Mislevy, R. (2018). *Sociocognitive foundations of educational measurement.* Routledge.

Molenaar, I., de Mooij, S., Azevedo, R., Bannert, M., Järvelä, S., & Gašević, D. (2023). Measuring self-regulated learning and the role of AI: Five years of research using multimodal multichannel data. *Computers in Human Behavior*, 139, Article 107540. https://doi.org/10.1016/j.chb.2022.107540

Morell, Z. (2017). *Introduction to the New York State next generation early learning standards.* https://www.nysed.gov/sites/default/files/introduction-to-the-nys-early-learning-standards.pdf

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science, 32*(1–2), 99–113. https://doi.org/10.1023/B:TRUC.0000021811.66966.1d

Moro, E., Frank, M. R., Pentland, A., Rutherford, A., Cebrian, M., & Rahwan, I. (2021). Universal resilience patterns in labor markets. *Nature Communications*, *12*, Article 1972. https://doi.org/10.1038/s41467-021-22086-3

Mumford, M. D., & Owens, W. A. (1987). Methodology review: Principles, procedures, and findings in the application of background data measures. *Applied Psychological Measurement, 11*(1), 1–31. https://doi.org/10.1177/014662168701100101

Murphy, S. C., Klieger, D. M., Borneman, M. J., & Kuncel, N. R. (2009). The predictive power of personal statements in admissions: A meta-analysis and cautionary tale. *College and University*, *84*(4), 83–86.

Narciss, S. (2004). The impact of informative tutoring feedback and self-efficacy on motivation and achievement in concept learning. *Experimental Psychology, 51*(3), 214–228. https://doi.org/10.1027/1618-3169.51.3.214

Narciss, S., Sosnovsky, S., Schnaubert, L., Andrès, E., Eichelmann, A., Goguadze, G., & Melis, E. (2014). Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers & Education*, *71*, 56–76. https://doi.org/10.1016/j.compedu.2013.09.011

National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures.* The National Academies Press. https://doi.org/10.17226/24783

National Academies of Sciences, Engineering, and Medicine. (2019). *Monitoring educational equity.* The National Academies Press. https://doi.org/10.17226/25389

National Association of Colleges and Employer. (2018). 2019 *NACE job outlook report.* https://www.odu.edu/content/dam/odu/offices/cmc/docs/nace/2019-nace-job-outlook-survey.pdf

National Association of Colleges and Employers. (2022). *NACE Job Outlook 2022*. https://www.naceweb.org/uploadedFiles/files/2022/resources/nace-job-outlook-2022.pdf

National Research Council. (1999a). *High stakes: Testing for tracking, promotion, and graduation*. The National Academies Press. https://doi.org/10.17226/6336

National Research Council. (1999b). *Myths and tradeoffs: The role of tests in undergraduate admissions.* The National Academies Press. https://doi.org/10.17226/9632

National Research Council. (2000). *How people learn: Brain, mind, experience, and school* (expanded ed.). The National Academies Press. https://doi.org/10.17226/9853

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* The National Academies Press. https://doi.org/10.17226/10019.

National Research Council. (2011). *Incentives and test-based accountability in education*. The National Academies Press. https://doi.org/10.17226/12521

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. The National Academies Press. https://doi.org/10.17226/13398.

Nesbit, J. C., Adesope, O. O., Liu, Q., & Ma, W. (2014, July). How effective are intelligent tutoring systems in computer science education? In *2014 IEEE 14th International Conference on Advanced Learning Technologies* (pp. 99 -103). IEEE. https://doi.org/10.1109/ICALT.2014.38

Nguyen, T. H., Han, H.-R., Kim, M. T., & Chan, K. S. (2014). An introduction to item response theory for patient-reported outcome. *Measurement*, *7*(1), 23–35. https://doi.org/10.1007/s40271-013-0041-0

Nickow, A., Oreopoulos, P., & Quan, V. (2020). *The impressive effects of tutoring on PreK-12 learning: A systematic review and meta-analysis of the experimental evidence* (NBER working paper no. 27476). National Bureau of Economic Research. https://doi.org/10.3386/w27476

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2017). Measuring non-cognitive predictors in high-stakes contexts: The effect of self-presentation on self-report instruments used in admission to higher education. *Personality and Individual Differences*, 106, 183–189. https://doi.org/10.1016/j.paid.2016.11.014

Noor, N., Beram, S., Yuet, F. K. C., Gengatharan, K., Syafiq, M., & Rasidi, M. S. M. (2023). Bias, halo effect and horn effect: A systematic literature review. *International Journal of Academic Research in Business & Social Sciences*, *13*(3), 1116–1140. https://doi.org/10.6007/IJARBSS/v13-i3/16733

Norville, V. (2022). States sketch 'portraits of a graduate.' *State Innovations, 27*(1). 1–4.

Novarese, M., & Di Giovinazzo, V. (2013). *Promptness and academic performance* (MPRA Paper No. 49746). Munich Personal RePEc Archive. https://mpra.ub.uni-muenchen.de/49746/

O'Dwyer, E., Sparks, J. R., & Nabors Oláh, L. (2023). Enacting a process for developing culturally relevant classroom assessments. *Applied Measurement in Education*, *36*(3), 286–303. https://doi.org/10.1080/08957347.2023.2214652

OECD. (n.d.). *Education & Skills Online Assessment*. https://www.oecd.org/skills/ESonline-assessment/aboutededucationskillsonline/

OECD. (2015). *Skills for social progress: The power of social and emotional skills.* OECD Publishing. https://doi.org/10.1787/9789264226159-en

OECD. (2019). An OECD Learning Framework 2030. In G. Bast, E. G. Carayannis, & D. F. J. Campbell (Eds.), The future of education and labor. *Arts, research, innovation and society* (pp. 23 -35). Springer. https://doi.org/10.1007/978-3-030-26068-2_3

OECD. (2021). *AI and the future of skills: Volume 1. Capabilities and assessments*. OECD Publishing. https://doi.org/10.1787/5ee71f34-en.

OECD. (2022). *Building the future of education*. OECD Publishing. https://web-archive.oecd.org/2022-11-30/618066-future-of-education-brochure.pdf

OECD. (2022). *PISA 2022 results*. https://www.oecd.org/publication/pisa-2022-results#pisa2022results

OECD (2023). *OECD skills outlook 2023: Skills for a resilient green and digital transition*. OECD Publishing. https://doi.org/10.1787/27452f29-en

Oh, I.-S., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: A meta-analysis. *Journal of Applied Psychology, 96*(4), 762–773. https://doi.org/10.1037/a0021832

O'Neil, H., Baker, E. L., Wainess, R., Chen, C., Mislevy, R., & Kyllonen, P. (2004). *Final report on plan for the assessment and evaluation of individual and team proficiencies developed by the DARWARS Environments.* Office of Navel Research; Defense Advanced Research Project Agency. https://apps.dtic.mil/sti/tr/pdf/ADA432802.pdf

OPM. (n.d.). *Other assessment methods*. OPM U.S. Office of Personnel Management. https://www.opm.gov/policy-data-oversight/assessment-and-selection/other-assessment-methods/

Ormerod, C. M., Malhortra, A., & Jafari, A. (2021). *Automated essay scoring using efficient transformer-based language models*. PsyArXiv. https://arxiv.org/pdf/2102.13136.pdf

Ortner, T. M., & Proyer, R. T. (2015). Objective personality tests. In T. M. Ortner & F. J. R. van de Vijver (Eds.), *Behavior-based assessment in psychology: Going beyond self-report in the personality, affective, motivation, and social domains* (pp. 133–149). Hogrefe.

Ortner, T. M., Proyer, R. T., & Kubinger, K. D. (2006). *Theorie und praxis objektiver personlichkeitstests* [Theory and practice of objective personality tests]. Verlag Hans Huber.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models* (No. 144). Sage.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89*(2), 187–207. https://doi.org/10.1037/0021-9010.89.2.187

Panadero, E. (2023). Toward a paradigm shift in feedback research: Five further steps influenced by self-regulated learning theory. *Educational Psychologist*, *58*(3), 193–204. https://doi.org/10.1080/00461520.2023.2223642

Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review, 35*, Article 100416. https://doi.org/10.1016/j.edurev.2021.100416

Panthier, C., & Gatinel, D. (2023). Success of ChatGPT, an AI language model, in taking the French language version of the European Board of Ophthalmology examination: A novel approach to medical knowledge assessment. Journal Français d'Ophtalmologie, *46*(7), 706–711. https://doi.org/10.1016/j.jfo.2023.05.006

Patrick, S. (2021). Transforming learning through competency-based education. *State Education Standard*, *21*(2), 23–29.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braubn, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Erlbaum.

Pavlik, P. I., Yudelson, M., & Koedinger, K. R. (2011). Using contextual factors analysis to explain transfer of least common multiple skills. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial intelligence in education: AIED 2011* (pp. 256 -263). Springer. https://doi.org/10.1007/978-3-642-21869-9_34

Phelps, R. P. (2019). Test frequency, stakes, and feedback in student achievement: A meta-analysis. *Evaluation Review, 43*(3–4), 111–151. https://doi.org/10.1177/0193841X19865628

Poropat, A. E. (2014). A meta-analysis of adult-rated child personality and academic performance in primary education. *British Journal of Educational Psychology*, *84*(2), 239–252. https://doi.org/10.1111/bjep.12019

Posso, A. (2016). Internet usage and educational outcomes among 15-year old Australian students. *International Journal of Communication, 10*, 3851–3876. https://ijoc.org/index.php/ijoc/article/view/5586/1742

Powers, D. E., & Fowles, M. E. (1997) The personal statement as an indicator of writing skill: A cautionary note. *Educational Assessment, 4*(1), 75–87. https://doi.org/10.1207/s15326977ea0401_3

Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., & Black, A. W. (2018). Style transfer through back-translation. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Vol. 1. Long Papers* (pp. 866–876). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1080

Pu, S., Converse, G., & Huang, Y. (2021, June). Deep Performance Factors Analysis for Knowledge Tracing. In *International Conference on Artificial Intelligence in Education* (pp. 331-341). Springer, Cham. DOI:10.1007/978-3-030-78292-4_27

Qian, Y., Tao, J., Suendermann-Oeft, D., Evanini, K., Ivanov, A. V., & Ramanarayanan, V. (2018a). *Computer-implemented systems and methods for speaker recognition using a neural network* (U.S. Patent 10,008,209). U.S. Patent Office and Trademark Office. https://ppubs.uspto.gov/pubwebapp/external.html?q=(10008209).pn.&db=USPAT&type=ids

Qian, Y., Tao, J., Suendermann-Oeft, D., Evanini, K., Ivanov, A. V., & Ramanarayanan, V. (2018b). Noise and metadata sensitive bottleneck features for improving speaker recognition with non-native speech input. In *Proceedings of INTERSPEECH 2016: 17th Annual Conference of the International Speech Communication Association* (pp. 3648–3652). https://doi.org/10.21437/Interspeech.2016-548

RAND. (2020). *RAND education assessment finder.* https://www.rand.org/education-and-labor/projects/assessments/tool.html

Randall, J. (2003). It ain't near 'bout fair: Re-envisioning the bias and sensitivity review process from a justice-oriented antiracist perspective. *Educational Assessment, 28*(2), 68–82. https://doi.org/10.1080/10627197.2023.2223924

Rees, A., (2021, December 27). The history of predicting the future. *Wired*. https://www.wired.com/story/history-predicting-future/

Rios, J. A., Ling, G., Pugh, R., Becker, D., & Bacall, A. (2020). Identifying critical 21st-century skills for workplace success: A content analysis of job advertisements. *Educational Researcher, 49*(2), 80–89. https://doi.org/10.3102/0013189X19890600

Roediger III, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17*(4), 382–395. https://doi.org/10.1037/a0026252

Roll, I., & Barhak-Rabinowitz, M. (2023). Measuring self-regulated learning using feedback and resources. In N. Foster & M. Piacentini (Eds.), *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/c93ac64e-en.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Rupp, A. A. (2018). Designing, evaluating, and deploying automated scoring systems with validity in mind: Methodological design decisions. *Applied. Measurement in Education, 31*(3), 191–214. https://doi.org/10.1080/08957347.2018.1464448

Russell, M. (2023). *Systemic racism and educational measurement: Confronting injustice in testing, assessment, and beyond*. Routledge.

Russell, S., & Norvig, P. (2022). *Artificial Intelligence: A modern approach* (4th ed.). Pearson.

Salgado, J. F., & Moscoso, S. (2019). Meta-analysis of interrater reliability of supervisory performance ratings: Effects of appraisal purpose, scale type, and range restriction. *Frontiers in Psychology, 10*, Article 2281. https://doi.org/10.3389/fpsyg.2019.02281

Salganik, M. J. (2023). Predicting the future of society. *Nature Human Behavior, 7*, 478–479. https://doi.org/10.1038/s41562-023-01535-7

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: a framework for constructing "intermediate constraint" questions and tasks for technology platforms. *The Journal of Technology, Learning and Assessment, 4*(6). https://ejournals.bc.edu/index.php/jtla/article/view/1653

Schmill, S. (2022, March 28). We are reinstating our SAT/ACT requirement for future admissions cycles in order to help us continue to build a diverse and talented MIT. *MIT Admissions.* https://mitadmissions.org/blogs/entry/we-are-reinstating-our-sat-act-requirement-for-future-admissions-cycles/#annotation-10

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.; pp. 307–353). American Council on Education; Praeger.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science*, *58*(8), 1438–1457. https://doi.org/10.1287/mnsc.1110.1509

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*(2), 331–354. https://doi.org/10.1007/BF02294343

Shafer, G. W., Viskupic, K., & Egger, A. E. (2023). Critical workforce skills for bachelor-level geoscientists: An analysis of geoscience job advertisements. *Geosphere*, *19*(2), 628–644. https://doi.org/10.1130/GES02581.1

Shen, T., Lei, T., Barzilay, R.,& Jaakkola, T. (2017). Style transfer from non-parallel text by cross-alignment. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *Advances in neural information processing systems 30 (NIPS 2017)* (pp. 1–12). Curran Associates. https://papers.nips.cc/paper_files/paper/2017/file/2d2c8394e31101a261abf1784302bf75-Paper.pdf

Shepard, L. A. (2017). Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The future of assessmen*t (pp. 279–303). Routledge. https://doi.org/10.4324/9781315086545-12

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*, 53–76. https://doi.org/10.1016/j.asw.2013.04.001

Schmitt, N, Keeney, J, Oswald, F. L., Pleskac, T. J., Billington, A. Q., Sinha, R., & Zorzie, M. (2009). Prediction of 4-year college student performance using cognitive and noncognitive predictors and the impact on demographic status of admitted students. *Journal of Applied Psychology*, *94*(6), 1479–1497. https://doi.org/10.1037/a0016810

Schrum, L., & Levin, B. B. (2013). Leadership for twenty-first-century schools and student achievement: Lessons learned from three exemplary cases. *International Journal of Leadership in Education*, *16*(4), 379–398. https://doi.org/10.1080/13603124.2013.767380

Schwartz, D. L., Tsang, J. M., & Blair, K. P. (2016). *The ABCs of how we learn: 26 scientifically proven approaches, how they work, and when to use them.* W.W. Norton & Company.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Sinatra, A. M., Robinson, R. L., Goldberg, B., & Goodwin, G. (2023). Impact of engaging with intelligent tutoring system lessons prior to class start. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *67*(1), 2262–2266. https://doi.org/10.1177/21695067231192709

Sinharay, S. (2023). Statistical methods for detection of test fraud on educational assessments. In R. J. Tierney, F. Rizvi, & K. Ercikan (Eds.) *International encyclopedia of education* (4th ed., pp. 298–307). Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.10030-2

Sinharay, S., & Johnson, M. S. (in press). Computation and accuracy evaluation of comparable scores on culturally responsive assessments. *Journal of Educational Measurement.*

Sinharay, S., & Johnson, M. S. (2012). Statistical modeling of automatically generated items. In M. J Gier & T. M. Haladyna (Eds.), *Automatic item generation* (pp. 183–195). Routledge.

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in educational assessment. *Educational Measurement: Issues and Practice, 39*(3), 100–105. https://doi.org/10.1111/emip.12377

Society for Industrial Organizational Psychology. (2018). *Principles for the validation and use of personnel selection procedures* (5th ed.). https://www.apa.org/ed/accreditation/personnel-selection-procedures.pdf

Soland, J., & Kuhfeld, M. (2021). Do response styles affect estimates of growth on social-emotional constructs? Evidence from four years of longitudinal survey scores. *Multivariate Behavioral Research*, *56*(6), 853–873. https://doi.org/10.1080/00273171.2020.1778440

Solano-Flores, G. (2019). Examining cultural responsiveness in large-scale assessment: The matrix of evidence for validity argumentation. *Frontiers in Education*, *4*, Article 2019.00043. https://doi.org/10.3389/feduc.2019.00043

Soto, C. J., Napolitano, C. M., Sewell, M. N., Yoon, H. J., & Roberts, B. W. (2022). An integrative framework for conceptualizing and assessing social, emotional, and behavioral skills: The BESSI. *Journal of Personality and Social Psychology, 123*(1), 192–222. https://doi.org/10.1037/pspp0000401

Sottilare, R. A., Baker, R. S., Graesser, A. C., & Lester, J. (2018). Special issue on the generalized intelligent framework for tutoring (GIFT): Creating a stable and flexible platform for innovations in AIED research. *International Journal of Artificial Intelligence and Education, 28*(1), 139–151. https://doi.org/10.1007/s40593-017-0149-9

Sparks, J. R., Lehman, B., & Zapata-Rivera, D. (2023). *Caring assessments: Challenges and opportunities* [Manuscript submitted for publication].

Stankov, L., Kleitman, S., & Jackson, S. A. (2015). Measures of the trait of confidence. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of personality and social psychological constructs* (pp. 158–189). Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-386915-9.00007-3

Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *Journal of Educational Psychology, 105*(4), 970–987. https://doi.org/10.1037/a0032447

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M., Snook, S. A., & Grigorenko, E. L. (2000). *Practical intelligence in everyday lif*e. Cambridge University Press.

Stecher, B. M., & Hamilton, L. S. (2014). *Measuring hard-to-measure student competencies: A research and development plan* (Research Report No. RR-863-WFHF). RAND Corporation. https://doi.org/10.7249/RR863

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*(3), 277–292. https://doi.org/10.1177/014662169301700308

Stowe, K., Ghosh, D., & Zhao, M. (2022). *Controlled language generation for language learning items.* arXiv. https://doi.org/10.48550/arXiv.2211.15731

Straub, L. M., Lin, E., Tremonte-Freydefont, L., & Schmid, P. C. (2023). Individuals' power determines how they respond to positive versus negative performance feedback. *European Journal of Social Psychology, 53*(7), 1402–1420. https://doi.org/10.1002/ejsp.2985

Su, R., Tay, L., Liao, H.-Y., Zhang, Q., & Rounds, J. (2019). Toward a dimensional model of vocational interests. *Journal of Applied Psychology, 104*(5), 690–714. https://doi.org/10.1037/apl0000373

Sykes, C. J. (1999, December 6). Soccer moms vs. standardized tests. *The New York Times.* https://www.nytimes.com/1999/12/06/opinion/soccer-moms-vs-standardized-tests.htm

Tang, R., Chuang, Y.-N., & Hu, X. (2023). *The science of detecting LLM-generated texts.* arXiv. https://doi.org/10.48550/arXiv.2303.07205

Tang, Z., & Kirman, B. (2023). Exploring curiosity in games: A framework and questionnaire study of player perspectives. *International Journal of Human-Computer Interaction*. Advance online publication. https://doi.org/10.1080/10447318.2024.2325171

Tannenbaum, R. J., & Kane, M. T. (2019). *Stakes in testing: Not a simple dichotomy but a profile of consequences that guides needed evidence of measurement quality* (Research Report No. RR-19-19). ETS. https://doi.org/10.1002/ets2.12255

Tenison, C., & Sparks, J. R. (2023). Combining cognitive theory and data driven approaches to examine students' search strategies in simulated digital environments. *Large-Scale Assessments in Education, 11*, Article 28. https://doi.org/10.1186/s40536-023-00164-w

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. Crown Publishers.

Turchin, D. (Host). (2023, March 6). Andi Mann, Sageable CEO and AIOps pioneer, discusses enterprise AI wins and the impact of automation on jobs [Audio podcast episode]. In *AI and the Future of Work. Apple Podcasts*. https://podcasts.apple.com/us/podcast/andi-mann-sageable-ceo-and-aiops-pioneer-discusses/id1476885647?i=1000602978601

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology, 9, 151–176. https://doi.org/10.1146/annurev-clinpsy-050212-185510*

U.S. Congress, Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions* (Report No. OTA-SET-519). U.S. Government Printing Office.

U.S. Department of Education, Office of Educational Technology. (2023). *Artificial intelligence and future of teaching and learning: Insights and recommendations.* https://www2.ed.gov/documents/ai-report/ai-report.pdf

U.S. Office of Personnel Management. (n.d.). *Situational judgment tests*. https://www.opm.gov/policy-data-oversight/assessment-and-selection/other-assessment-methods/situational-judgment-tests/

van der Linden, W. J. (2005). *Linear models for optimal test design*. Springer Science + Business Media. https://doi.org/10.1007/0-387-29054-0

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. Springer. https://doi.org/10.1007/978-0-387-85461-8

van de Vijver, F. J. R.,& He, J. (2016). Bias assessment and prevention in noncognitive outcome measures in context assessments. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning: International perspectives* (pp. 229–253). Springer. https://doi.org/10.1007/978-3-319-45357-6_9

van de Vijver, F., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. H. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment.* Taylor & Francis. https://doi.org/10.4324/9781410611758

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. (2007). When are tutorial dialogues more effective than reading?. *Cognitive science, 31*(1), 3-62.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52*(1), 8.

von Davier, M., & Bezirhan, U. (2023). A robust method for detecting item misfit in large-scale assessments. *Educational and Psychological Measurement, 83*(4), 740-765.

von Davier, M., Tyack, L., & Khorramdel, L. (2023). Scoring graphical responses in TIMSS 2019 using artificial neural networks. *Educational and Psychological Measurement, 83*(3), 556 -585. https://doi.org/10.1177/00131644221098021

Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, 104, Article 106189. https://doi.org/10.1016/j.chb.2019.106189

Wainer, H. (1987). *The first four millennia of mental testing: From ancient China to the computer age* (Research Report No. RR-87-34). ETS. https://doi.org/10.1002/j.2330-8516.1987.tb00238.x

Wainer, H., & Thissen, D. (2001). True score theory: The traditional method. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 35-84). Routledge. https://doi.org/10.4324/9781410604729

Walker, M. E., Olivera-Aguilar, M., Lehman, B., Laitusis, C., Guzman-Orth, D., & Gholson, M. (2023). *Culturally responsive assessment: provisional principles* (Research Report No. RR-23-11). ETS. https://doi.org/10.1002/ets2.12374

Walkington, C., & Bernacki, M. L. (2020). Appraising research on personalized learning: Definitions, theoretical alignment, advancements, and future directions. *Journal of Research on Technology in Education, 52*(3), 235–252. https://doi.org/10.1080/15391523.2020.1747757

Wang, J., Jou, M., Lv, Y., & Huang, C. C. (2018). An investigation on teaching performances of model-based flipping classroom for physics supported by modern teaching technologies. *Computers in Human Behavior, 84*, 36-48.

Weinberger, C. J. (2014). The increasing complementarity between cognitive and social skills. *Review of Economics and Statistics, 96*(5), 849 -861. https://doi.org/10.1162/REST_a_00449

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied psychological measurement*, *41*(2), 115-129. https://doi.org/10.1177/0146621616676791

Weiss, S., Wilhelm, O., & Kyllonen, P. (2021). An improved taxonomy of creativity measures based on salient task attributes. *Psychology of Aesthetics, Creativity, and the Arts. Advance online publication.* https://doi.org/10.1037/aca0000434

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wilkie, D. (2023, December 21). *Employers say students aren't learning soft skills in college*. https://www.shrm.org/topics-tools/news/employee-relations/employers-say-students-arent-learning-soft-skills-college

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*(4), 52-61.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*, Article 3087. https://doi.org/10.3389/fpsyg.2019.03087

Wolcott, M. D., Lobczowski, N. G., Zeeman, J. M., & McLaughlin, J. E. (2020). Situational judgment test validity: an exploratory model of the participant response process using cognitive and think-aloud interviews. *BMC medical education, 20*, 1-12. https://doi.org/10.1186/s12909-020-02410-z

World Economic Forum (2021). *Building a Common Language for Skills at Work: A Global Taxonomy*. https://www3.weforum.org/docs/WEF_Skills_Taxonomy_2021.pdf

World Economic Forum (2022). Catalysing Education 4.0 Investing in the Future of Learning for a Human-Centric Recovery Insight Report. https://www3.weforum.org/docs/WEF_Catalysing_Education_4.0_2022.pdf

World Economic Forum (2023). Defining Education 4.0: A Taxonomy for the Future of Learning. https://www3.weforum.org/docs/WEF_Defining_Education_4.0_2023.pdf

Xuan, Q., Cheung, A., & Sun, D. (2022). The effectiveness of formative assessment for enhancing reading achievement in K-12 classrooms: A meta-analysis. *Frontiers in Psychology*, 13, Article 990196. https://doi.org/10.3389/fpsyg.2022.990196

Yang, Z., Hu, Z., Dyer, C., Xing, E. P., & Berg-Kirkpatrick, T. (2018). Unsupervised text style transfer using language models as discriminators. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31 (NeurIPS 2018*) (pp. 7287–7289). Curran Associates. https://papers.neurips.cc/paper_files/paper/2018/hash/398475c83b47075e8897a083e97eb9f0-Abstract.html

Yeung, C. (2019). *Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory.* arXiv. https://doi.org/10.48550/arXiv.1904.11738.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences, 112*(4), 1036 -1040. https://doi.org/10.1073/pnas.1418680112

Zapata-Rivera, D., & Forsyth, C. M. (2022). Learner modeling in conversation-based assessment. In R. A. Sottilare, & J. Schwarz (Eds.), *Adaptive instructional systems: International Conference on Human-Computer Interaction. HCII 2022* (pp. 73–83). Springer. https://doi.org/10.1007/978-3-031-05887-5_6

Zapata-Rivera, D., & Hu, X. (2022). Assessment in intelligent tutoring systems SWOT analysis. In A. M. Sinatra, A. C. Graesser, X. Hu, G. Goodwin, & V. Rus (Eds.), *Design recommendations for intelligent tutoring systems: Vol. 10. Strengths, weaknesses, opportunities and threats (SWOT) analysis of intelligent tutoring systems* (pp. 83–90). US Army Combat Capabilities Development Command – Soldier Center. https://gifttutoring.org/attachments/download/4751/Vol%2010_DesignRecommendationsforITSs_SWOTAnalysisofITSs.pdf#page=87

Zapata-Rivera, D., Lehman, B., & Sparks, J. R. (2020). Learner modeling in the context of caring assessments. In R. A. Sottilare & J. Schwarz (Eds.), *Adaptive Instructional Systems: Second International Conference (AIS) 2020* (pp. 422–431). Springer.

Zhan, J., Her, Y. W., Hu, T., & Du, C. (2018). Integrating Data Analytics into the Undergraduate Accounting Curriculum. *Business Education Innovation Journal, 10*(2), 169 -178. http://www.beijournal.com/images/V10N2_draft_5.pdf

Zhang, Z. (2012). Microsoft kinect sensor and its effect. *IEEE Multimedia*, *19*(2), 4-10. https://doi.org/10.1109/MMUL.2012.24

Zu, J., & Choi, I. (2023a, April 12–15). *Utilizing deep language models to predict item difficulty of language proficiency tests* [Paper presentation].The annual meeting of National Council on Measurement in Education, Chicago, IL, United States.

Zu, J., & Choi, I. (2023b, July 25–28). P*redicting the psychometric properties of automatically generated items* [Paper presentation]. International Meeting of the Psychometric Society, College Park, MD, United States.

Zu, J., & Kyllonen, P. C. (2020). Nominal response model is useful for scoring multiple-choice situational judgment tests. *Organizational Research Methods, 23*(2), 342–366. https://doi.org/10.1177/1094428118812669